# The role of automated agents in collective actions on social media: Individuals, interactions and networks




Linda Li

Linacre College

University of Oxford

*Thesis submitted in partial fulfilment of the requirements for DPhil in Social Data Science in the Oxford Internet Institute*

Word count: 69,396

Supervisors: Prof Helen Margetts and Prof Balazs Vedres

Michaelmas Term 2024

# Abstract

Social media profoundly influences personal lives and civil discourse, and social bots have emerged as significant actors in promoting and disrupting civic discussions and activism on it. This dissertation addresses the gap in the lack of empirical studies examining bots' impact on political communication and online activism.

Focusing on two social movements cases, Extinction Rebellion (XR) and Black Lives Matter (BLM), the research investigates how automated accounts influence political discourse and collective actions. It explores three dimensions of bot impact: individual behaviour, social network, and the discourse shaped by both bots and humans during protests.

The study begins by developing a robust method for identifying social bots in social media platforms. Using XR as a case study, this study then finds that social bots and human users create information cascades around climate protests, significantly altering human sentiment and posting behaviour.

At the meso level, the research analyses how bots affect human-to-human connections within online networks during BLM protests. It identifies communities or 'botnets' of social bots that cluster based on differing motivations, which contribute to a decline in human cohesiveness within social networks. Users with different attitudes towards BLM are affected differently by these bots.

Building on these findings, the dissertation investigates the mechanisms behind social bots' effects on human discourse, which is a key component of online activism. It refines existing theoretical frameworks to account for the hybrid nature of online activism involving both human and automated agents and reveals how social bot strategies can create, constrain, or reshape the discursive opportunities for activists.

Overall, this research enhances our understanding of social bots in online activism, offering significant theoretical, empirical, and methodological contributions. It has important implications for activists and policymakers and suggests various future directions in the study of social bots and online activism.

I may not have gone where I intended to go, but I think I have ended up where I needed to be.

– *The Long Dark Tea-Time of the Soul, Douglas Adams*

獨有南山桂花發，飛來飛去襲人裾

——《長安古意》盧照鄰

# Acknowledgements

After writing countless statements of purpose, personal statements and cover letters where I was explicitly told by people *not* to start with this cliché, I finally have a place to say it: when I was 15, I read Michel Foucault's *Discipline and Punish*.

It was a spring afternoon, just before dusk and about to rain, and the air tasted like wet soil. I sat under the grey, overclouded sky of southern China and told myself that I would major in sociology for my undergraduate degree. I did. Twelve years later, I am still here. Even better, I have also learned to code.

It has definitely been a long journey, and for this, my first and foremost gratitude goes to my supervisors, Prof Helen Margetts and Prof Balazs Vedres. I am grateful to Helen for encouraging and mentoring me to be an independent social scientist, to push the boundaries of human knowledge, and to think and write with clarity, precision, and coherence. I am grateful to Balazs for his guidance in developing my scholarly professionalism, conducting rigorous quantitative empirical studies, and going through the publication of my first academic journal article step by step. I am indebted to both of you for your continuous support in many aspects of my DPhil journey, especially during the particularly turbulent days. Without the guidance and mentoring of either of you, I would not have become who I am today.

Next, I hope to express my sincere gratitude to other scholars who have supported me during my DPhil study. I would like to thank Dr Orsolya Vásárhelyi (Corvinus University), co-author of one of my papers, for her enormous help and support. I am grateful to Orsi for providing constructive suggestions and contributions and for introducing me to the basics of surviving as an early-career scholar in academia. I would also like to thank Dr Charles Rahal and Dr Michael Biggs. Charlie and Michael encouraged me to pursue a DPhil when I was an MSc student at Oxford and offered continuous support and constructive advice throughout my research and my entire DPhil journey. Special thanks to Charlie for being an amazing MSc supervisor and to Michael for being my assessor and for his invaluable comments during those milestones, together with my other milestone assessors, Dr Scott Hale, Dr Luc Rocher, Prof Joss Wright and Prof Sandra González-Bailón.

I am indebted to the multiple grants and scholarship providers who funded this DPhil and associated research expenses, including the Shirley Scholarship, the OII - Dieter Schwarz Foundation, the Funds for Women Graduates, the Great Britain-China Educational Trust, the MacArthur Foundation, as well as Linacre College

and the Oxford Internet Institute. Without this generous support, it would have been impossible for me to complete this DPhil.

I am grateful to the helpful, friendly, and professional administrative team at OII. Thank you, Christine Bunyan, Duncan Panssey, Laura Mayard, David Pepper, Caroline Watson, Karen Mead, and Sarah Moore, for streamlining the administrative procedures at OII so seamlessly and for continuously and sincerely taking care of the students here, including me.

I would also like to thank those at and outside OII who had constructive discussions with me about my research and my career in academia in general: Dr Jonathan Bright, Prof Man Yee Kan, Prof Victoria Nash, Dr Mary Sanford, Dr Hubert Au, Dr Patrick Gildersleve, Dr Mathis Ebbinghaus, Dr Arun Frey, Dr Hanbo Wu, Dr Agnes Horvat. This list may not be comprehensive, but I am truly grateful for the insightful conversations I had with these wonderful people.

Apart from those who helped me complete my dissertation, I am thankful to some excellent collaborators (and often mentors as well) in my side research programmes. I am grateful to Prof David Stark (Columbia University), Prof Ekaterina Hertog, Dr Naomi Muggleton (University of Warwick), and the research team at the Leverhulme Centre for Demographic Science. During the process of working together with them, I have learned an incredible amount and am extremely grateful.

I would like to thank the course convenors who offered me teaching positions, including Dr Bernie Hogan, Dr Scott Hale, Dr Fabian Stephany, Dr Carl Frey, Prof Andrew Przybylski, and Prof Ekaterina Hertog (again), as well as the Oxford-Stanford Exchange Programme and the Oxford-Williams Exchange Programme. I would also like to thank my fellow TAs and RAs, Dragos Gourduza, Christine Sheldon, Freya Hope, Michael Collyer, and Dr Tiffany Ge Wang, for making it a pleasure to work with them.

I am grateful to Dr Zhiyuan Yu and Prof Ronggui Huang at Fudan University, Prof Daniel Garvey (emeritus) at Prescott College, Keran W. and Soheil R. at UNESCAP, Dr Bernie Hogan at OII (again), as well as Simin D. and Yifei C. from the Oxford MSc/MPhil in Sociology cohort for encouraging me to pursue graduate study in the very beginning.

The OII DPhil student cohort has been a truly warm, welcoming, and friendly one, and I am extremely thankful to my friends here. I would like to first thank my social data science DPhil cohort—Anna George, Marcel Schliebs, Elizaveta Chernenko, Dr Hannah Bailey, Tarah Wheeler, and Alexandra Pavliuc. I particularly owe deep gratitude to Anna, Marcel, and Lisa for your support in many ways during the probably most difficult spring and summer of my DPhil. I would also like to thank Calvin Yixiang Cheng, Dr Julia Slupska, and Dr Cailean Osborne

for being truly excellent roommates; it was indeed a pleasant experience sharing a flat with you. Additionally, I want to thank Diyi Liu and Yixin Chen for your mental support and the heartwarming food you shared.

I would like to thank my family and extended family for the significant role they have played in shaping who I am today. Their influence has undeniably impacted the past ten thousand days of my life. I am grateful for what they have devoted to me and fully aware of the profound influence they have had on my life.

Finally, I feel particularly obliged to acknowledge some names. The last few years often felt like navigating a boat through a foggy, turbulent ocean, guided by the faint glimmer of a distant star. Along the way, I owe much to the persistent presence of these individuals. Thank you, Xingxing C., Yuwei Z., Jingfeng W., Yaowen D., Jieping C., Yumeng J., Moting J., Letian W., Xinyi L., Dr Weirong G., and Junchao T. Thank you, Zayna B., Dilys C., and the college, department, and university' s welfare service teams. Thank you to my online tabletop RPG group and the few remarkable networks of friends. Thank you, Dr Esteban G. M.

Thank you, past self, for the decision you made twelve years ago and for what you have stayed committed to over the years.

# Contents

**List of Figures** **ix**

**List of Tables** **x**

**1 Introduction** **1**

1.1 Background and research questions . . . . . 1
1.2 Thesis overview . . . . . 8
1.3 Case selection . . . . . 10
1.4 Ethical considerations . . . . . 12

**2 Literature Review** **14**

2.1 Social bots . . . . . 16
2.2 Activism and social media . . . . . 33
2.3 Platforms . . . . . 44
2.4 Context of protest cases . . . . . 48

**3 Methodology** **58**

3.1 Data collection . . . . . 59
3.2 Bot identification . . . . . 62
3.3 Text categorisation using ChatGPT . . . . . 68
3.4 Topic modelling . . . . . 69
3.5 Approach to identifying potential causal relationships . . . . . 72
3.6 Conclusion . . . . . 75

**4 The impact of social bots on human behaviour in protest communication** **77**

4.1 Introduction . . . . . 78
4.2 Results . . . . . 82
4.3 Limitations . . . . . 90
4.4 Discussion . . . . . 92
4.5 Data and methods . . . . . 95

**5 The impact of social bots on online protest networks 103**
5.1 Introduction . . . 104
5.2 Findings . . . 111
5.3 Discussions . . . 120
5.4 Data and methods . . . 124

**6 The strategies of social bots in shaping online activism discourse 134**
6.1 Introduction . . . 135
6.2 Theoretical framework and research gaps . . . 139
6.3 Methodology . . . 150
6.4 Findings . . . 153
6.5 Discussions and conclusions . . . 177

**7 Discussions and conclusions 183**
7.1 Findings . . . 184
7.2 Contributions of the dissertation . . . 190
7.3 Practical implications . . . 195
7.4 Limitations . . . 198
7.5 Future research directions . . . 201
7.6 Conclusions . . . 205

**Appendices**

**A Supplementary information for Chapter 4 208**

**B Supplementary information for Chapter 5 251**

**C Supplementary information for Chapter 6 276**

**D Authorship statement 283**

**Bibliography 286**

# List of Figures


3.1 Sample output of Botometer bot classifications outcomes. . . . . 63
3.2 Proportion of users classified as 'bots' (red) and 'humans' (blue) in the dataset, with bot thresholds ranging from 0 to 1.0. . . . . 65

4.1 Information flow of humans and bots. . . . . 83
4.2 Number of tweets posted by bots and humans in a cascade. . . . . 85
4.3 Two examples of Twitter timelines. . . . . 87
4.4 Predicted change resulting from a bot interaction. . . . . 88

5.1 Two example of changes in ego network clustering, differed by bot interaction frequency. . . . . 116
5.2 Point estimators of bot interaction's impact on changes in individual and community level connections. . . . . 117
5.3 Line estimators of bot interaction's impact on changes in connections, moderated by support categories to BLM. . . . . 119

6.1 Breakdown of bot-generated messages by content. . . . . 155

# List of Tables


2.1 A review of bot strategies in existing literature . . . . . 28

3.1 Keywords collected for data collection. . . . . 62
3.2 List of the training datasets. . . . . 66
3.3 Pearson Correlation between the support level values of Twitter users labelled by ChatGPT and manual coders. . . . . 69

4.1 VECM models testing the relationship between the amount and sentiment of bot and human communication. . . . . 86

5.1 Summary of network metrics and characteristics for each community cluster. . . . . 113

6.1 DOS definition and sub-dimensions . . . . . 138
6.2 Datasets used for computer-assisted content analysis. . . . . 152
6.3 Bot strategies grouped by DOS dimensions . . . . . 154
6.4 Distribution of Resonance Categories by Bot Type in BLM and XR Datasets . . . . . 162
6.5 Distribution of Legitimacy Categories in Bot Activities for BLM and XR Datasets . . . . . 168
6.6 Overview of bot activities in the DOS of the two cases . . . . . 175

# 1 Introduction

## Contents


**1.1 Background and research questions . . . . . 1**
**1.2 Thesis overview . . . . . 8**
**1.3 Case selection . . . . . 10**
1.3.1 Extinction Rebellion (XR) . . . . . 11
1.3.2 Black Lives Matter (BLM) . . . . . 11
**1.4 Ethical considerations . . . . . 12**


## 1.1 Background and research questions

As a critical component of the public sphere, communication processes in contemporary times are heavily shaped and facilitated by online technologies[1]. We exist in an era of 'deep mediatisation' where social media platforms such as X (formerly Twitter), Facebook, YouTube, and Instagram saturate both our personal lives and civic discussions[2]. The prevalence of social media has significantly reduced the cost of social connection between individuals while promoting information diffusion. Those changes shaped both personal connection and the public sphere[3, 4].

Apart from personal communication and public debates, social media has also revolutionised online collective actions or online activism. Social media platforms provide a powerful tool for activists to amplify their messages, mobilise supporters,

and engage in collective action[5, 6, 7]. Collective action—organised efforts to advocate for social or political change—is a natural extension of these communication processes[8]. From environmental movements[9] to racial justice protests[5], activists have made use of the connective power of social media to reach a global audience, often bypassing traditional media gatekeepers.

However, this digital public sphere is subject to disruption. Alongside positive civic engagement, the ecology of social media faces challenges such as divisive content, polarisation, and external attempts to silence discussions[10, 11]. These disruptive forces not only affect individual interactions but also influence larger, collective efforts to mobilise around social causes, demonstrating the increasingly complex relationship between social media, communication, and activism.

One of the potential powers to influence the social media public is automated accounts, also known as social bots. These two concepts, automated accounts or social bots, refer to computer programs that possess the ability to complete automated tasks online, such as engaging with humans on social media platforms[12]. As automation becomes more prevalent in the last decade, they are becoming a significant element that influences communication and discourse on social media. In many social media conversations surrounding political events, such as the 2016 U.S. presidential election and the Brexit referendum[13, 14], past research have observed orchestrated bot activities. Social bots can disseminate a substantial volume of messages at an incredibly low cost and engage with other users promptly and rapidly, surpassing the capabilities of human labour[15].

These characteristics enable social bots to both promote and disrupt civic discussions, which are important to democratic societies. On the one hand, it has become increasingly common for activists on social media to use automated accounts to facilitate protest mobilisation. Traditional media outlets also employ automation to amplify the reach of their messages. For instance, an X bot can be designed to respond to users who mention their state in a tweet by providing them with contact details for their state's senators, along with a suggestion to urge the senators to support gun-control proposals[16]. Such applications demonstrate the

beneficial use of social bots in aiding civic engagement. On the other hand, there are bots controlled by malicious agents that mimic human behaviour and automatically generate a large number of online messages[17]. Many of these messages are fake, hyper-partisan, or inflammatory, possessing the potential to disrupt the healthy ecosystem of online media[18, 13]. Whether beneficial or malicious, social bots possess unique abilities that allow them to exert a significant and distinct influence on online discourse and communication networks, ultimately shaping political processes. Consequently, it is both epistemologically and practically important for us to conduct systematic analyses of social bot activities and their impact.

While there has been research on the definition, taxonomy, and detection of social bots[19, 20], there are still significant gaps in our understanding of their impact, particularly in relation to political and civic discourse, and online activism. Existing research about social bots has focused on identifying bots or describing their behaviour, yet little work has empirically measured how bots influence human interactions online or the outcomes of political events. This is especially true in the context of activism, such as protests, where the role of social bots remains under-explored.

Research on civic discourse and online activism has traditionally concentrated on the role of social media in shaping social movements along three key dimensions: communication, movement organisation, and mobilisation[5, 6, 7]. Given the prevalence of bots on social media, these automated accounts are likely influencing all three dimensions, yet their specific impact remains under-examined. However, it is worth noting that earlier studies often removed automated agents from datasets when analysing online political discourse[21, 22], thus neglecting an important factor in understanding the full scope of political conversations online. More recent research has begun to investigate how bots influence public opinion and mobilisation during political events, but these studies remain limited in scope[23, 18, 24]. These studies inspected bots' impact in multiple dimensions, such as their influence on public opinion and protest mobilisation networks. However, we still lack a study

that systemically and comprehensively researches social bots' impact on all aspects of the dynamic between political events and social media.

Furthermore, the few pieces of existing research of social bot impact focuses on institutionalised political events like elections and their related discussions on social media, while only a limited number of studies examined the role of bots in protests and protest discussions online. However, collective action is also an essential component of modern-day democracy[25, 26]. It is a democratic process that continues throughout the electoral cycle[5], with evolving dynamics, making it equally important to study in its own right. It is also more and more common for protests to use social media as an important tool for information diffusion and protest mobilisation[27, 28]. Therefore, it is essential to enrich the existing body of literature by adding a systematic study focusing on bots' impact on online activism and discourse related to them.

This doctoral project seeks to fill these research gaps by focusing on social bots' role in collective actions on social media. By doing so, it first makes two key contributions to the existing literature on social media agents, with a specific focus on social bots.

First, this research deepens our understanding of social bots by examining their strategies and the specific ways they influence online discourse. While extensive research has been conducted on identifying bots and describing their behaviour patterns [17, 12], there has been considerably less attention given to exploring their strategic use and the impact they have on human behaviour. By addressing these gaps, this study contributes to a more nuanced understanding of how social bots operate within social media environments and affect the dynamics of online interactions.

Second, this research broadens the conversation to encompass the societal effects of automated agents, especially in shaping online activism. Much of the existing research on social media discourse and civic activism has predominantly centred on human actors. However, as social bots become increasingly sophisticated, they are playing a critical role in influencing social media interactions. These automated

accounts can now effectively mimic human behaviour, thereby impacting the way information is disseminated and engaged with in the digital landscape [15]. This research not only enhances our understanding of social bots as active participants in online discourse, but also advances theoretical perspectives on automated agents-or even more broadly, artificial intelligence (AI)-and their broader societal impact. By investigating the influence of social bots, this study enriches our comprehension of automated agents, or even AIs' role in communication networks, social network analysis, and media discourse studies.

This study also contributes to the broader research area of collective action on social media in two key aspects. It first deepens our understanding of how social media environments impact activism. Social media has become a essential tool for modern-day activism, functioning as both a mean of mobilisation and a space where activists shape public discourse. Protest movements rely on social media not only to coordinate actions but also to raise awareness, amplify messages, and influence public opinion, with the ultimate aim of impacting policy outcomes[29, 30, 31]. By examining the role of automated agents such as social bots in these processes, this research enriches our understanding of how social media affects activism, especially social bots. This exploration is particularly relevant in an age when online activism is a critical component of protest movements, underscoring the importance of studying the new dynamics introduced by automated agents.

This research makes the second contribution in this area by addressing the sociological complexity of collective action, particularly the challenges involved in mobilising and sustaining engagement in protest movements. According to Rogers et al.[32], collective actions are paradoxical, involving 'the effort to marshal many people to perform behaviours that impose a net cost on each individual who complies and provide negligible collective benefit unless performed by a large number of individuals.' Reaching the critical mass necessary for collective benefit requires overcoming significant obstacles, such as motivating participation, maintaining momentum, and resisting disruptions. While existing theories of collective action

explore these dynamics, there is a growing need for frameworks that account for the disruptive potential of social bots.

Social bots, as automated agents, introduce a new layer of complexity to these sociological challenges. They can amplify protest messages, boost engagement[33], or alternatively, create barriers to mobilisation by spreading disinformation, undermining trust, or disrupting communication channels[34]. Despite their increasing presence in digital activism, the specific ways in which bots influence mobilisation and collective action remain under explored. This dissertation fills this gap by systematically investigating how social bots interact with protest movements and examining whether they facilitate or hinder mobilisation efforts. In doing so, the research provides important insights into how these automated agents influence the dynamics of collective action. Moreover, by incorporating social bots into the analysis of collective action, this study contributes to the development of new theoretical models that reflect the evolving digital landscape, updating traditional sociological frameworks to account for the impact of both social media and automation on human mobilisation.

Furthermore, this research makes methodological contributions by developing a bot identification method tailored to the research context. While various bot identification methods and tools are available, such as the open-source tool Botometer[35], they are reported to be less effective in the field of social science research, particularly in the fields of political communication and political sociology[36, 37]. Through a systematic review of previous studies on bot identification, this research compares different methods in the context of online activism, and proposes a practical and replicable procedure for political bot detection. This methodological contribution enhances the existing knowledge on bot identification and benefits social science research that requires accurate bot detection as a prerequisite.

Finally, beyond its academic contributions, this research will also offer valuable social and practical implications in the field of social bots and online activism. By clarifying the mechanisms and extent of social bots' impact, this study can inform policy recommendations and guide platform-level regulations. Furthermore, it aims

to raise awareness among stakeholders and the general public regarding important issues such as regulations of AI, and the responsible use of data for social good. The insights gained from this research not only contribute to a deeper understanding of the dynamics of online activism but also empower activists and the broader community by equipping them with knowledge to navigate the challenges posed by automated agents in the digital landscape.

To make those contributions, this dissertation project aims to understand the role of automated accounts in collective actions on social media through three dimensions: impact on individual-level behaviour, impact on social networks and the strategies bot employ in online interactions on social media. It starts by identifying social bots on social media, then clarifying their impact, and finally understanding bot activities that lead to this impact. Due to the advancements in AI seen in recent years, this study have limited its scope to social bot activities between 2018 and 2022. This period reflects a time when existing research reported consistent traits and strategies of social bots. This research chose not to discuss in detail developments after 2022, such as the emergence of more advanced automated agents driven by large language models (LLMs), which significantly altered the nature of bots. However, my research remains highly relevant, as it focuses on the foundational traits and strategies of social bots during a period of relative consistency, and provides valuable insights into their impact on social media and political communication. The analysis is done with a mixture of social data science and qualitative methods. The research questions (RQs) are as follows, which are also elaborated in the 'thesis overview' section:

- **Central RQ**: How do automated accounts affect online collective actions and related discussions on social media?
- **RQ1 (Answered in Chapter 4)**: How do they influence human communication behaviour in protest-related discourse?
- **RQ2 (Answered in Chapter 5)**: How do they impact the network structure of online social networks in protests?

- **RQ3 (Answered in Chapter 6)**: What are the strategies social bots use to affect discourse in online activism?

## 1.2 Thesis overview

The dissertation is organised in the following structure, as an attempt to answer the overarching research questions stated above. Based on an overview of the context and background of the research, the project starts by stating research questions and summarising the cases studied for the project (Chapter 1). Then, it provides a comprehensive literature review on the research topic (Chapter 2). Chapter 3 introduces the methodological bases of the entire dissertation project, such as data collection, natural language processing (NLP), and bot identification.

The project then proceeds to address the research questions. Chapters 4 and 5 start with a quantitative measurement of the scale and extent of bots' potential impact on protest-related discourse and social media networks. Additionally, this project supports the proposed relationships using causal inference methods, though other factors may contribute to those effect. To achieve this, it initially explores whether exposure to social bots influences individual behaviour in online communication related to protests. Building upon this inquiry, this study also investigates whether such micro-level influence is linked with shifts in online opinions. Furthermore, since exposure to social bots is often short-term or sporadic, this research examines whether they have a lasting impact on human behaviour into the future. All these aspects are examined in Chapter 4 through the empirical case of multiple consecutive waves of Extinction Rebellion (XR) protests that occurred between November and December 2019, addressing RQ1.

In Chapter 5, the project progresses to understanding the impact of bots at the meso-level: how bots shape the online social network, to address RQ2. This is important because social networks reflect the structure of human collaborations and protest organisation. This study uses the case of Black Lives Matter (BLM) and focuses its analysis on the protests that occurred after George Floyd's death in late May to early June 2020. Similar to the examination of individual-level

behaviour, this chapter also explores the impact of automation on both the peak of protests and the long term.

After showing an association and potential causal link between bot activities and changes in human behaviour, Chapter 6 offers a qualitative analysis of how social bot activities may lead to these changes. It examines the specific activities of social bots and their impact on human behaviour. Using computationally assisted content analysis (CACA) methods, Chapter 6 explores the interactions between humans and bots embedded in online discourse during the two protest cases. The central question addressed is: What are the strategies social bots use to affect discourse in online activism? (RQ3) This focus is important because social bots influence protest activists through text and discourse, which are integral to how activists' agendas are disseminated and received by the public, thereby affecting the outcome of protests. This study analysed three dimensions of social bot impact on online activism discourse, building on the discursive opportunity structure (DOS) theory: visibility, resonance, and legitimacy. For each dimension, it details bot strategies and their effects on the DOS. This analysis contributes to previous research by illustrating and conceptualising the specific ways in which social bots shape the effectiveness and public perception of activist discourse.

In the final chapter (Chapter 7), this research provides an overview of the dissertation, contributions to research fields, practical implications, and limitations of the project. It also proposes future directions for research based on the insights gained from this dissertation.

Finally, this thesis follows the format of an integrated thesis, with the empirical chapters (sections of Chapter 3 and Chapters 4, 5, and 6) based on articles prepared for journal publication. These chapters, therefore, include a) a preamble that summaries how they address the RQs of this dissertation and b) a short declaration of authorship and publication status. The appendices contain further supplementary information for those chapters, as well as signed declarations for any co-authorship.

## 1.3 Case selection

The focus of the dissertation project is on examining the impact of bots on collective actions occurring on social media. To address this research question, this project analyses two collective action cases with political objectives: Extinction Rebellion (XR) and Black Lives Matter (BLM). The selection of these cases is based on two rationales, including the nature of the protests and the feasibility of conducting empirical research. Both XR and BLM are large-scale protests that have gained international attention, and activists involved in these movements have been highly active on various social media platforms, particularly Twitter. The significant scale of these events has facilitated extensive discussions and the generation of abundant data capturing the complete landscape of information flow and protest mobilisation processes in international collective action events[38, 39, 40]. Furthermore, both cases exhibit clear protest goals supported mostly by the left wing. However, the online discourse surrounding these goals is highly polarised, resulting in active streams of anti-protest information flow and counter-protests, such as the #AllLivesMatter movement[41]. This polarised nature of the protest-related discourse on Twitter provides an opportunity to analyse bot activity spanning the ideological spectrum, serving both left- and right-wing, pro- and anti-protest, as well as political and non-political purposes. Additionally, existing studies and news reports have indicated the presence of automation in discussions related to climate change and racial issues on social media[42, 43]. For instance, during the Paris Summit in 2016, a significant proportion of messages on Twitter were found to be generated by social bots[42], and bots have been observed spreading conspiracy theories about the BLM movement. These observations further support the relevance of studying the role of bots in these specific topics.

Below is a brief overview of the two cases. More detailed review of past work concerning the two cases can be seen in Chapter 2.

### 1.3.1 Extinction Rebellion (XR)

XR, which stands for Extinction Rebellion, is a grassroots global environmental social movement that originated in the UK in October 2018 and has remained active since then. According to their own descriptions, the movement employs non-violent disobedience tactics to draw attention to ecological destruction and climate change, urging governments to take action[44]. XR activists claimed that they employ various tactics, including occupying public spaces, intentionally causing disruption, lobbying parliament, and willingly subjecting themselves to arrest[45]. Since winter 2019, XR activists have organised a series of non-violent disruptive events across the UK and in numerous cities spanning 35 countries worldwide. These actions primarily target capital cities, based on the study of successful campaigns in the past. Throughout these movements, XR activists heavily rely on social media platforms such as Facebook, Instagram, and Twitter to reach out to the public and disseminate their message.

### 1.3.2 Black Lives Matter (BLM)

The BLM movement can be traced back to the #BlackLivesMatter hashtag that originated in July 2013. As an activist movement, it started in 2014 in Ferguson, Missouri with the street demonstrations following the death of Michael Brown on August 9, 2014. It was originated to raise awareness on police killings of unarmed African American citizens and has been ongoing since then to fight against racism, discrimination, and racial inequality[39]. BLM activists relied on social media, especially Twitter, as a critical way to spread their messages and call for actions nationally and internationally. Offline collective actions events mostly included mass protests, demonstrations and other contentious events.

For this dissertation project, I focus on the outburst of protests events happening during the period of late May to early June 2020 after George Floyd's death in May 27, 2020. The death of George Floyd on May 25, 2020, in Minneapolis, Minnesota caused the Black Lives Matter (BLM) movement to have a significant resurgence. George Floyd, an African American man, died while in police custody

when a police officer knelt on his neck for nearly nine minutes during an arrest. The incident was captured on video and widely shared on social media, leading to widespread outrage and calls for justice, as well as widespread protests and demonstrations across the United States[46].

## 1.4 Ethical considerations

This research involves the collection of social media data using methods such as digital ethnography. The data collected includes sensitive personal information like political preferences, which raises distinct ethical considerations. I have obtained the Central University Research Ethics Committee (CUREC)' s approval for the project, where I detailed the ethical considerations and measures taken to address these challenges (CUREC reference: SSH_OII_CIA_21_109). This study also adheres to relevant guidance documents, especially Oxford CUREC' s Best Practice Guidance (BPG) for Ethnography and Internet-Mediated Research[47]. More detailed discussions of the ethical considerations are provided in the relevant chapters of this dissertation.

The major ethical considerations fall into the categories of informed consent, discussion of sensitive issues, and data security. For Chapter 4 and Chapter 5, this study collects and analyses social media posts, which involves ethical issues related to informed consent, data privacy, and security. This study removes all Twitter profiles set to private from the adopted dataset and excludes them from the analysis. Additionally, since bots are typically one-off accounts, the collection of their data often involves using information that has already been deleted. Given that automated accounts are not human entities, the current ethical standards do not require informed consent for data collected from them. However, with ongoing developments in AI, this may evolve in the future. For now, this study adheres to established guidelines while ensuring that all ethical considerations for human participants are rigorously observed.

For Chapter 6, this study uses an unobtrusive approach that involves mostly observing online communities and conducting content analysis. This research is

conducted primarily on publicly available data from public social media platforms, which is generally considered 'publicly available data' according to the BPG. This research also adheres to other guidance from the BPG for internet-mediated research, including the considerations and risks associated with publishing tweets and the implications for the researcher.

In summary, this dissertation aims to provide a systematic answer to the question of how social bots impact online political communication and activism. Starting with an extensive literature review, it proceeds with an in-depth investigation of two major protest cases. It first designs a robust, replicable research method for studying social bots. Then, it investigates social bots' potential impact on human behaviour and online social networks, followed by a qualitative investigation of how social bot activities lead to these changes. By doing so, this study contributes to research areas such as automated agents, social media, and social movement studies, theoretically, empirically, and methodologically.

# 2
# Literature Review

## Contents


**2.1 Social bots . . . . . 16**
2.1.1 Definitions . . . . . 17
2.1.2 Taxonomy . . . . . 20
2.1.3 Identification . . . . . 22
2.1.4 Behaviour and strategies . . . . . 27
2.1.5 Source and operators . . . . . 31
**2.2 Activism and social media . . . . . 33**
2.2.1 Information dissemination . . . . . 34
2.2.2 Network and organisation . . . . . 38
2.2.3 Mobilisation tactics . . . . . 41
**2.3 Platforms . . . . . 44**
2.3.1 Platforms and bots . . . . . 44
2.3.2 Platforms and activism . . . . . 46
**2.4 Context of protest cases . . . . . 48**
2.4.1 Extinction Rebellion (XR) . . . . . 48
2.4.2 Black Lives Matter (BLM) . . . . . 52


This chapter provides a comprehensive literature review addressing the central research question (RQ): How do social bots impact online activism? To explore this question effectively, the review is structured to cover several critical areas. The chapter begins with a discussion of the key actors involved in the dynamics of online activism: social bots, activists, and the platforms they operate on. These three actors form the foundation upon which the subsequent analysis is built, and their

interaction shapes the broader landscape of digital activism.

First, the chapter examines existing literature on social bots, focusing on five main areas: their definitions, taxonomies, identification methods, operational strategies on social media, and origins. Understanding these aspects is important because they lay the groundwork for assessing how social bots function and influence online environments. This section establishes the technical and operational dimensions of bots as key actors on the digital stage.

Next, the review examines how social media shapes modern protests by addressing three key aspects: information dissemination, network formation, and protest mobilisation. This section summarises past work on digital activism, especially in contexts where social bots have played a role, and highlights how activists make use of digital platforms in their strategies. Understanding these dynamics is essential for framing the interaction between bots and human-driven activism.

Furthermore, though social media platforms are not the central focus of this research, their design, features and rules significantly influence both the operation of social bots and the activities of online activists. This chapter will therefore examine the infrastructure and policies of platforms, which act as both enablers and regulators of bot behaviour and activism. Additionally, this section will justify the dissertation's focus on analysing bot activities on X (formerly Twitter), given its prominence as a site for both bot deployment and activist engagement.

Finally, the chapter provides background on the two major protest movements studied in this dissertation: Extinction Rebellion (XR) and Black Lives Matter (BLM), and how the aforementioned 'actors'-bots, activists and platforms-come together in those cases. These cases are representative of contemporary contentious political movements that extensively make use of social media, exhibit significant traces of automation and address pressing environmental and racial issues which are prominent topics in elections, news media, and civic activism. This discussion contextualises the interaction between bots, activists, and platforms in real-world movements, providing concrete examples of how these actors play out on the global stage of digital activism.

Each section of the chapter summaries existing research, highlights gaps, and indicates which chapters will address these gaps in more detail. By mapping the interactions of bots, activists, and platforms, this review not only provides a thorough overview of the relevant literature but also lays the foundation for a deeper exploration in subsequent chapters, identifying areas where further research is essential to understanding the impact of social bots on online activism.

## 2.1 Social bots

This dissertation focuses on social bots, and this chapter first reviews existing studies on five main areas of social bot literature: their definitions, taxonomy, identification methods, operators' strategies on social media, and their origins. Despite extensive research on social bots, there are still gaps in conceptualisation, such as the lack of a universally accepted definition of social bots. Furthermore, social bots are studied in several disciplines, including sociology, communication studies, and computer science. Despite their prevalence in these distinct fields, there is an absence of meaningful dialogue and collaboration between bot studies across disciplines. Consequently, this section aims to synthesise past research on social bots from diverse disciplines, constructing a comprehensive overview of existing studies. Drawing insights from this collective body of work, this chapter subsequently proposes potential directions for future research informed by this interdisciplinary review.

As mentioned in Chapter 1, this dissertation project focuses on social bots and relevant past work within the time frame of this dissertation project (2018 to 2022). As such, this section specifically covers past work on social bots before recent advancements of social bots, such as those driven by Large Language Models (LLMs) after 2022. Although these advances have significantly increased the sophistication of automated accounts and introduced new strategies and behaviours, they do not make earlier research obsolete.

Although we acknowledge that the literature reviewed here reflects a specific period in the evolution of social bots, this does not undermine the validity of

this project. This section provides an overview of the fundamental definitions, taxonomy, and behavioural patterns of social bots, which remain largely consistent with contemporary automated systems in 2024. While LLM-driven bots demonstrate greater linguistic sophistication, their core strategic objectives, amplification techniques, and behavioural dynamics build upon pre-existing bot frameworks. Furthermore, understanding pre-LLM bot strategies is essential for contextualising newer developments, as many of these advancements refine rather than entirely replace prior methodologies. By establishing a foundation in past bot behaviours, this section still contributes to a broader understanding of the evolving role of social bots in digital discourse.

### 2.1.1 Definitions

Despite significant scholarly attention[19, 12] to the concept of 'social bots' over the past decade, the term remains relatively ambiguous. Researchers and journalists often either omit a precise definition or present varying and ambiguous interpretations of the concept. Additionally, past studies frequently use the term 'social bots' interchangeably with related concepts such as 'bots,' 'automated agents,' 'robots,' and even AI without adequately distinguishing between them. This lack of consistency often leads to misunderstandings and confusion, hindering a deeper and more systematic understanding of the topic. Therefore, this review begins by elucidating the usage of these concepts in existing literature to establish a foundation for this research.

One issue that arises prior to further discussion of social bots is the question of agency—whether these automated accounts are merely tools or if they possess the capacity to act independently. While this literature review does not extensively explore the issue of bot agency, the issue of bot agency directly influences how one defines social bots and shapes our understanding of their role in online environments. If bots are considered to have agency, it impacts how one interprets their behaviour, activities, and potential influence, as well as how one uses the term

in research and discourse. Therefore, this review briefly addresses the concept of agency to provide context for the subsequent literature review.

Only a limited body of literature specifically addresses the question of whether social bots, or automated accounts, possess agency. A few studies have explicitly discussed the philosophical or ecological aspects of whether bots can be considered to have agency[48, 20] or, on the contrary, lack it[49]. However, the majority of studies tend to make implicit assumptions about this question, shaping their descriptions of bot definitions and activities accordingly (e.g. 'social bots are hidden actors with a political agenda'[20]). While not a direct focus of this literature review, it will reference existing literature while preserving the original expressions used by authors to convey their perspectives. This approach ensures that the perspective of each study is respected, while the review maintains focus on clarifying the broader conceptual landscape surrounding social bots.

As the name suggests, social bots constitute a sub-category of 'bots,' which is a shortened form of 'software robots.' However, the term is frequently used interchangeably with the concept of 'automated agents.' The definition of the broader concept of software robots predominantly comes from the field of computer science, automation, and AI ethics. According to Duh et al.[50], automated agents are described as systems situated within and integrated into an environment, capable of sensing that environment and taking actions over time to pursue their designated objectives and influence future events. Furthermore, the term 'bots,' as implied by its name, specifically refers to automated agents used in digital spaces primarily created to automate routine virtual tasks, instead of physical robots with tangible bodies designed for performing physical tasks[51].

Since the concept was initially coined in the 1990s[52], its meaning in prior research has undergone several transformations and has incorporated definitions from multiple disciplines. When the term was first introduced, it was used to refer to a wide range of software systems that perform jobs with some extent of automation. This encompassed daemons and certain programs designed to engage

in pre-written text-based interactions with human users, often referred to as 'chatbots'[51, 53]. Subsequently, in the early 21st century, the term gained widespread usage in the field of information security, describing computers controlled by or compromised due to malware[54]. With the rise of online social media platforms like Twitter, automated accounts found their place to disseminate content and interact with human users. Consequently, earlier works, particularly within the field of social studies, also began to designate these accounts as bots'[55]. They are alternatively termed 'Sybils' in Internet security literature[56].

This dissertation project specifically focuses on the last category of bots, commonly known as 'social bots.' Ferrara et al. proposed a widely accepted definition of social bots as a type of computer program capable of automatically generating content and engaging with humans on the Internet[12]. Other researchers have employed the term in slightly different ways. According to Grimme et al.[20], social bots are automated software agents with the capacity to act autonomously on the web. Meanwhile, prior work in the field of social sciences has developed their definition of the term based on the communicative or socio-political implications of at least some social bots. They have suggested that a essential characteristic of social bots is their ability to mimic human behaviour in order to interact with other users[55], while others argue that social bots must possess a political agenda.

This research adopts the definition of 'social bots' as outlined by Ferrara et al.[12] for analysis because it offers the broadest and most inclusive framework. This broader definition is preferable to more specific ones, such as those focusing solely on 'human-mimicking agents'[55], which may not encompass all relevant automated activities for this study. For instance, research on online political communication and protests has demonstrated that automated agents perform a variety of tasks beyond mimicking human behaviour[23, 57]. Limiting the definition to only those agents that mimic human behaviour might miss other significant forms of automated activity. Adopting a broader definition allows for a more comprehensive examination of all relevant automated activities, including those that may impact human behaviour and political processes[58, 59].

Based on this broad definition, one also needs a comprehensive taxonomy to classify the various types of social bots effectively. This is important because a well-defined taxonomy helps in systematically categorising different forms of automated activity, which can vary significantly in their functions and impacts. To address this need, this chapter will review existing typologies of social bots in the next section. By examining these typologies, one can better understand the diverse ways social bots operate and influence online environments. Building on the discussions in this section and the next ('Taxonomy'), Chapter 6 of this dissertation explores how the definition and categories of social bot activities are constructed in the context of online activism, considering insights from both identification methods and online discourse. Addressing this research gap lays the groundwork for exploring RQ3 (bot strategies), as a thorough understanding of the nature and types of bots is essential for comprehending their strategies and implications.

### 2.1.2 Taxonomy

Existing research has developed three major categories of taxonomies for bots, which this section briefly reviews and summarises. It is important to note that past studies often employ the terms 'bots' or 'social bots' according to their individual definitions. Here, this study adheres to the concept of Ferrara et al.[12], which defines social bots as automated agents on social media, and it evaluates all typologies of automated agents that fall within this definition. For coherency, this chapter will consistently use the term 'social bots' throughout this review.

For the first category, Stieglitz et al.[55] and Ferrara et al.[12] introduced a two-axis taxonomy for social bots. According to Stieglitz[55], bots can be categorised along two dimensions: their level of malice towards the online community and the degree to which they mimic human behaviour. 'Malice' refers to the extent to which their actions either bolster or disrupt existing community rules, while 'human-mimicking' refers to the degree to which social bots aim to create the impression that they are human when interacting with human users. According to this taxonomy, most social bots involved in online political communication are

both malicious and human-mimicking, as they pose as human users and disseminate misinformation or inflammatory content under this guise[18]. Malicious, non-human-mimicking bots encompass those that post spam messages on social media. Finally, in the category of 'non-human' and 'non-malicious' bots, an example is Wikipedia bots[60, 61], which explicitly identify themselves as automated and actively contribute to the maintenance of community regulations.

The second type of taxonomy categorises social bots based on the designers' intent or social function. As proposed by Grimme et al.[20], the term 'social bot' serves as a broader concept that encompasses various types of (semi-) automated agents. These agents are specifically created to achieve distinct purposes through one- or multi-sided communications on online media. Oentaryo et al.[62] roughly classified social bots into three categories: broadcast bots (who help disseminate information), consumption bots (who provide updates of services or compile information from elsewhere), and spam bots. Monaco et al. further classified bots into the following categories:

a) Accounts on social media specifically designed to interact with humans. They refer to this category as 'social bots' under their definition.

b) Chatbots, bots that are specifically designed to converse with humans.

c) Service bots designed to cover routine automated tasks, such as bots deployed by news media to automatically publish and disseminate news stories.

d) Spambots, or bots whose main purpose is to send out a massive amount of messages, often used for malicious purposes such as political misinformation or money fraud.

e) Cyborgs, or so-called 'bot-assisted humans' or 'human-assisted bots.' The definition of this category is relatively ambiguous, as there are no set standards of the extent to which automation needs to be involved in a social bot to make it a cyborg.

For the third type of typologies, several prior studies aimed to establish a more comprehensive framework for bot classification with multiple dimensions. It's important to note that these frameworks are typically designed for the concept of 'bots' in general, rather than solely focusing on social bots. For example, Duh et al.[50]

argued that to characterise a bot, it is essential to describe its environment, sensing, possible actions, drives, and decision-making. Maus[63] and Monaco et al.[64] proposed a similar framework encompassing the following aspects: transparency (whether the bot discloses its bot status), level of automation, coordination with other bots, interaction and passivity, and intention.

It's important to note that existing taxonomies aim to encompass a wide range of social bots with diverse purposes and motivations. These taxonomies often adopt a generalised approach to traverse the disciplinary gap among bot studies from different fields. Although this approach contributes to a holistic comprehension of diverse social bots, existing taxonomies lack the granularity required to account for all differences in specific contexts. For instance, social bots supporting a protest and its counter-protest may exhibit distinct behaviours and strategies, yet they are often categorised broadly as 'political bots' in current taxonomies[57, 23].

In the empirical chapters, especially chapter 6, this study will partially address this gap by giving a classification of social bot activities in the context of online activism. This also addresses RQ3 (bot strategies), as understanding these distinctions is essential for analysing their impact and strategies effectively.

### 2.1.3 Identification

Another significant portion of the literature on social bots has primarily focused on the identification of these bots, especially on X (formerly Twitter). This emphasis arises due to the widespread presence of social bots in online communities and, as mentioned earlier, the fact that some of these bots are both human-mimicking and malicious[12]. They pretend to be non-automated users to enhance their credibility[65], facilitating the dissemination of misinformation and controversial content[24], thereby exacerbating conflicts[18]. To assess the extent of their impact and develop countermeasures, it is important to identify the presence of human-mimicking social bots. While it is true that social bots are a concern on various social media platforms, including Reddit and Facebook[66, 67], the issue is most

pervasive on X[17]. Consequently, the majority of existing bot identification literature has predominantly focused on Twitter bots as well. Therefore, for the purpose of this review, this chapter restricts the scope to bot identification on X.

The majority of bot identification research comes from the field of computer sciences, especially machine learning and information security. At the same time, bot studies from social sciences often adopt existing methods devised by the former. However, some social science bot researchers also developed their own methods to align with specific research objectives. This section will provide a detailed exploration of these methodologies.

Four categories of bot identification methods have emerged: profile feature-based bot detection, social network-based bot detection, crowdsourcing-based bot detection, and honeypot-based bot detection.

The first category, feature-based bot detection methods, focuses on identifying behavioural patterns that distinguish bot-like from human-like behaviour[68]. The specific features considered in these studies can vary, ranging from simple criteria like daily posting frequency to complex models encompassing a variety of online behavioural patterns[19]. However, the following categories of features are commonly taken into account[69, 19]:

a) Temporal behaviour, which includes characteristics such as rapidly posting a large number of messages in a short time frame or being active at times when human users are typically not active.

b) Text posted by the account, including elements like the Uniform Resource Locators (URLs) shared.

c) Profile information, such as the use of default profile settings.

d) Friends and social network traits, where social bots are known to follow and retweet each other, forming dense botnets.

Many of these studies choose to combine two or more categories of features mentioned above in their analyses. Furthermore, feature-based bot detection methods often rely on supervised machine learning classification algorithms. These algorithms typically use a training dataset consisting of labelled bot and human

user accounts, selecting a set of features that differentiate between the two user types. Subsequently, they classify a Twitter user based on these features and provide a prediction of the likelihood and degree to which they are automated.

Several publicly available feature-based bot detection tools are commonly adopted by social scientists for research purposes. One of the most well-known tools is Botometer[35] (formerly BotOrNot), which relies on machine learning and was initially released on 1 May 2014, with the latest update in September 2020. The developers of Botometer explain that it is designed to calculate a score where low scores indicate likely human accounts, while high scores suggest likely fully automated accounts[35]. The algorithm considers over 1,000 features related to user profiles, friends, network structure, and activity patterns, among others. Botometer has been widely used by social scientists for identifying Twitter bots in their research[70, 71, 24, 43]. Additionally, there are two other available tools (though their status may have changed) that claimed to be capable of identifying human-mimicking bots and were previously used by social scientists: 'heavy automation' and Tweetbotornot[72, 73].

In addition to these public tools, existing studies on social bots, especially those within political contexts and within the field of social sciences, have developed their own sets of features for identifying social bots. These studies have argued that existing public bot identification tools have false positive issues[72], leading to the misclassification of many human users as bots and vice versa. Moreover, existing bot detection tools used training sets that included all categories of social bots, whereas studies of social bots may focus on a specific type of social bot within a particular category. For example, many studies on social bots have concentrated on bot activities during election campaigns and online social movements[23, 74, 14]. These studies used training and testing datasets of social bots identified and labelled in online communications and developed their own sets of features for bot detection. For instance, studies of Gilets Jaune Movements and the Catalan Referendum and[23, 18] all devised their own bot detection methods.

The second category of bot identification methods comprises network-based (graph-based) or activity-based approaches. This method was developed because bots, often organised as a 'supervised network of zombie agents'[13], frequently engage in coordinated behaviour. They are known to follow each other on Twitter, repost each other's content, or replicate the same content across various social media platforms[13, 75, 76]. These behavioural patterns make network-based social bot detection feasible because they are easily distinguishable. Prior research has demonstrated three activity-based approaches to address the issue of bot identification:

a) Flow-based anomaly detection, which begins by capturing bot activities characterised by abnormal peaks in quantified net-flow[77].

b) Social network analysis methods, encompassing techniques like community detection, centrality measures, and clustering measurements[13, 78].

c) Identifying pivotal nodes, often exemplified as a recurring group of users simultaneously retweeting each other[77, 79].

These approaches have been employed collectively in previous studies to effectively identify social bot activities across various contexts, including the Brexit referendum[13], the Arab Spring[80], peer-to-peer (P2P) business promotion[81], and spambots responsible for disseminating phishing links[82, 81] .

The third and fourth categories of bot detection methods involve varying degrees of manual labour and subjective judgement, instead of focusing solely on quantifiable characteristics of potential social bots. In the third category, crowdsourcing-based bot detection relies on the collective efforts of individual human users who identify accounts exhibiting behaviour that resemble of automated accounts[12]. Most social media platforms permit their users to report messages or accounts that they believe violate site regulations, including malicious social bots. In prior studies, Wang et al.[83] adopted crowdsourcing methods for bot detection on Facebook, and Kouvela et al.[84] proposed a publicly available bot detection web service with user feedback functionality to enable crowdsourced bot detection. Crowdsourcing has also been widely employed in data annotation tasks for social bot detection[85].

The fourth category of bot detection methods employs honeypot accounts, which are accounts deliberately designed to attract automated agents for interaction[86, 87, 88]. After such interactions, the profile information of these social bots is recorded. This approach has been extensively used for two primary purposes: first, to generate datasets of social bots for further machine-learning-based bot detection tools, and second, to detect targeted cyberattacks so as to preventing social network infiltration.

Each of these four methods has its own advantages and drawbacks. Feature-based, machine-learning-based bot detection methods are efficient, scalable, and widely applicable across various contexts. However, as mentioned earlier, they are known for their false positive issues[72]. Furthermore, these algorithms and models are content and language-specific, as the behaviour patterns of bots heavily depend on the context they operate in[89]. Additionally, human-mimicking bots are constantly evolving, leading to changes in their behaviour patterns, which can render existing models relatively quickly outdated[19].

In contrast, social network-based, honeypot-based, and crowdsourcing-based bot detection methods appear to directly address these drawbacks. Coordinated activity patterns provide a clearer signal of bot activities. Simultaneously, human users tend to be quite sensitive to deviations from their own behaviour when it comes to automated patterns, making crowdsourced bot detection more accurate than feature-based detection. However, these methods can be computationally demanding[90, 91] or require substantial human labour input, making them less scalable options.

Finally, it is important to note that the emergence of derivative AI technologies, like ChatGPT, has made it increasingly challenging to identify bots due to their rapid advancement in mimicking human behaviour[92]. As large language models become more widespread and easily accessible through APIs, new social bots can exhibit extremely human-like behaviours, making it significantly more difficult to differentiate between individual bots and human users, even for experts[92].

However, during the time frame of this research's analysis (2018 to 2020), feature-based bot detection remains a valuable tool. Even in the post-proliferation of derivative AI, the detection of strategies and coordinated behaviours of social bots is still feasible[93]. Furthermore, bot detection at the strategy level continues to hold meaningful implications for defending against malicious activities carried out by social bot actors[37].

Based on these discussions, the bot identification pipeline this dissertation adopts is described and discussed in detail in Chapter 3, which serves as a foundation to address RQ1 (bot impact on behaviour), RQ2 (bot impact on network) and RQ3(bot strategies). The following section of this literature review reviews prior research on the strategies and behaviours of social bots, particularly those of human-mimicking and malicious bots.

### 2.1.4 Behaviour and strategies

The majority of studies examining social bot strategies have primarily concentrated on malicious, human-mimicking bots, with relatively few pieces of research investigating the roles and activities of non-malicious social bots. Additionally, the activities of non-malicious or non-human-mimicking bots tend to be straightforward and can be succinctly described by their names or intended functions[2]. For example, broadcast bots used by news media simply automate the dissemination of up-to-date news reports[62]. In contrast, the activities and strategies of malicious, human-mimicking bots tend to be more intricate and nuanced. Therefore, this section focuses on summarising past research concerning these types of social bots. Table 2.1 shows a overview of those strategies.

Research on malicious, human-mimicking social bots and their strategies predominantly centres on two aspects: their behavioural patterns and the content they attempt to disseminate. Although these bots often impersonate human users, there is a limited body of research on their profiles or the specific types of users they choose to emulate. This gap in research may stem from ethical concerns, as studying such aspects invariably involves disclosing information about specific

**Table 2.1:** A review of bot strategies in existing literature

| Strategy category | Strategy type | Description |
|---|---|---|
| Behaviour | Astroturfing | Bots post similar messages simultaneously to generate campaigns or information cascades. |
| | Mutual Following and Retweeting | Bots follow and retweet each other to reinforce astroturfing effects and increase message visibility. |
| | Targeting Influential/ Susceptible Users | Bots strategically reach out to influential or susceptible users to maximize the impact of their messages. |
| | Creating Divisive Narratives | Bots tailor divisive narratives for different user groups to enhance propaganda effectiveness. |
| Content | Posting Emotion-Evoking Content | Bots post inflammatory, emotional, or controversial content, especially around political events. |
| | Spreading Misinformation | Bots create and disseminate fake news, conspiracy theories, and disinformation. |
| | Amplifying Human Messages | Bots boost the visibility of human-posted messages, often negative or partisan, by retweeting and propagating them. |
| | Supporting Left-Wing Protest Movements | Bots support progressive or protest causes by amplifying relevant activities or impersonating political figures. |

accounts. Nevertheless, this study reviews existing studies concerning these two facets of bot strategies accordingly. Given the scope of this dissertation, its primary focus is on reviewing bot strategies in political contexts, though it may touch upon their behaviour in other contexts as well. Because of this specific context, existing studies predominantly come from the fields of political science, communication studies, and sociology.

The first dimension of bot strategy research centres on their behavioural patterns, including when and how they post messages on Twitter and which users they choose to follow. Past studies have predominantly reported two types of strategies.

First, bots engage in 'astroturfing' to promote the messages they intend to spread. According to Keller et al.[75], astroturfing involves a strategy where a group of agents simultaneously post similar messages to generate campaigns or information cascades. Astroturfing by bots has been observed in various time-sensitive political events on social media, such as the Brexit Referendum or the US presidential election[13, 42]. In these instances, automated accounts generate and post highly similar messages that are often hyperpartisan, extreme in viewpoint,

or involve fake news and rumours[13, 18, 74]. Astroturfing behaviour is known to create information cascades, promoting the messages bots aim to disseminate and creating the illusion of a lively discussion[75, 94]. During this process, bots reciprocally follow and retweet each other to reinforce the effect of astroturfing. For example, social bots may follow or contribute to trending political hashtags, as observed in cases like the protests against Turkey's censorship, the Ukraine conflict (in 2016), and elections in Brazil[94, 95]. Consequently, susceptible human users may be drawn to these activities in online conversations, potentially being exposed to and influenced by them.

Second, social bots reach out to non-automated users to increase the exposure and influence of their messages. They tend to strategically select their targets to maximise the impact of their messages. For instance, Stella et al.[18] suggested that during the Catalan referendum for independence, social bots on the periphery of the online communication network actively established connections with influential accounts. They also used violent content to inundate some users and intentionally sought to increase their exposure to negative and controversial material. Beyond highly influential individuals, social bots were found to selectively target individuals who were susceptible to disinformation or who displayed partisan alignment with them[24, 69, 71]. In the context of climate change communication, approximately 25% of the users they reached out to established connections with them[42]. Additionally, they were also observed to create divisive narratives tailored to different categories of targeted users to make their propaganda more successful[96].

The second aspect of bot strategy research is about the content that bot accounts post on social media. In essence, it deals with what social bots publish in their online interactions. The findings of these studies are highly context-dependent, as the specific content of bot-posted messages ultimately relies on the context of their activities. Nevertheless, there are common traits and strategies that can be summarised. Content posted by malicious, human-mimicking social bots can generally be classified into three major categories.

First, social bots often post and self-promote messages that deliberately aim to evoke strong human emotions. This is particularly prevalent in discussions related to political events[13, 97, 70]. In these discussions, social bots have been observed orchestrating efforts to post and disseminate inflammatory, highly controversial, negative, and/or hyperpartisan messages[18]. These messages frequently exhibit extreme attitudes and sentiments[98]. For example, in COVID-19 discussions, bots promoted misinformation in an attempt to 'create moral panic' around COVID-19 vaccinations[99]. In political communication related to Japan's general election, bots were reported to attack certain ethnic groups using extreme slang terms to fuel racist sentiments[97].

Second, bots have also been found to disseminate misinformation, fake news, or conspiracy theories. In such cases, they deliberately create and propagate fabricated claims with the hope of convincing human users to believe in false information. These activities have been observed across various political and non-political events, including the 2016 U.S. presidential and 2018 midterm elections[100, 101], the Brazil presidential elections[95], and The UN Climate Change Conference (COP25)[42], among others.

Third, social bots aim to amplify existing messages posted by human users, with a particular focus on messages from political actors. They achieve this by targeting and boosting the visibility of certain messages, thereby extending their reach and influence. For example, bots often retweet and propagate content that aligns with specific sentiments, including negative or pessimistic comments. This amplification process not only increases the dissemination of these messages but also enhances their impact and visibility within public discourse[102].

Most of those studies of bot strategies are based on isolated cases, which limits our ability to develop a comprehensive understanding. A systematic conceptualisation and taxonomy of bot strategies are important because they allow us to identify patterns, compare cases across different contexts, and generalise findings.

Furthermore, while existing research on social bots has predominantly focused on right-wing or conservative campaigns, evidence indicates that left-leaning or

progressive groups also use similar tactics, which remain under-researched. For example, during the 2013 political crisis in Venezuela, radical opposition groups employed bots to impersonate political leaders and agencies[33]. In Brazil, bots were used in 2016 to amplify anti-corruption protests on Twitter[103]. Despite these instances, there is a notable gap in understanding the role of 'good bots'—which could support pro-protest activities and positively influence discourse.

The next section (Protest and social media) will include a review of current research on bot strategies and their effects in the context of social movements. Chapters 4 and 5 will investigate these strategies within the two case studies and assess their impacts, addressing RQ1 (bot impact on behaviour) and RQ2 (bot impact on protests). Chapter 6, when addressing RQ3 (bot strategies), will explore the emergence of new strategies in these protests and provide a systematic conceptualisation of bot activities in contentious contexts. During the process, the chapter will pay particular attention to the under-researched role of pro-protest bots as well.

### 2.1.5 Source and operators

While not a primary focus of this study, a pertinent question to understand the behaviour of social bots is, 'Where do social bots come from?' Answers to this question are vital for a comprehensive analysis of bot behaviour, as knowing their origins can shed light on their design, purpose, and potential biases, which in turn helps interpret their impact and activities more accurately. However, only a limited corpus of research has examined the origins of social bots, and they are mostly from the field of information security and general Internet studies. This is because many social bots, especially those mimicking human behaviour, choose not to disclose their source. Despite this, existing studies provide insights into potential origins, which are briefly summarised in this section.

Early investigations of bot origins focused on open-source code management platforms like GitHub and GitLab, initiated by Kollyani et al.'s[104] analysis of bot codes shared on GitHub. Examining over 4,000 publicly available Twitter bot code repositories created from 2008 to 2016, they observed a predominance of bot

code originating in the United States. Most of these repositories were deemed 'side projects' or 'fun experiments' by experienced programmers. The bots fulfilled diverse functions, ranging from disseminating news to providing automatically generated updates for a service. Furthermore, about 1% of those bots serve political purposes. For example, some of those political bots used public data to generate a regularly updated stream of tweets or raise awareness on political topics.

Subsequent research by Wessel et al. extended the survey period to 2020, noting persistent patterns but with a shift in the geographical location of bot codes, particularly with countries like Japan emerging[105]. Millimaggi and Daniel[106] conducted a detailed analysis of social bot code patterns, identifying 28 communication-specific patterns with potential for abusive behaviour. Their research shed light on how social bots could disrupt online public communication. These patterns included:

> 'Disclosing sensitive facts, denigrating, being grossly offensive, being indecent or obscene, being threatening, making false allegations, deceiving users, spamming, spreading misinformation, mimicking interest, cloning profiles, and invading spaces.'[106]

However, there is a notable gap between open-source code repositories and the speculated existence of sophisticated, malicious social bots. While the creators of open-source bot code are generally identifiable, the users of these codes—who may employ them for harmful purposes—often remain unknown[104].

Some studies suggest a market for social bots and infrastructure components. For example, Frischlich et al. examined online sales spaces in both the 'clearnet' and the 'darknet,' revealing market trends[107]. However, it's important to note that they acknowledge the predominantly supportive nature of most social bot providers, who do not sell full-fledged, very sophisticated social bot services. Instead, they only offer specific modules of automation, such as generating fake profiles or functioning as fake followers.

Due to the covert nature of social bot operations, research on their origin often focuses on specific cases. Identifying social bots is increasingly challenging due

to their growth in sophistication. Instances become public knowledge through platform regulations or failed fraudulent acts, such as Twitter's discovery that 2,752 accounts active during the 2016 U.S. presidential election were managed by the Internet Research Agency (IRA), a company specialising in online influence operations based in St. Petersburg, Russia[108].

## 2.2 Activism and social media

The rapid development of social media has permeated various aspects of our lives, influencing critical social, political, and economic processes. Concurrently, there has been a notable surge in social unrest, from the Arab Spring and the Occupy Street movement in the early 2010s to the more contemporary ones such as Fridays for Future, the #MeToo movement, and new waves of Black Lives Matter activism. This parallel evolution has given rise to a wealth of literature exploring the dynamic interplay between social movements and social media over the past decade. Social media emerges as a distinct political infrastructure, particularly for social movements. Its prevalence brings about changes in different aspects of social movements and even alters the nature of such movements.

Past research has examined the distinct impact of social media on social movements, emphasising three key dimensions: communication, movement organisation and network, and movement mobilisation. At the same time, bots, the research focus of this dissertation, have become a significant phenomenon on social media. In the context of social movements, bots may potentially affect all aspects of the dynamic between protest and social media because of their prevalence and their distinct features (such as the rapid spread of a large number of messages). They can impact protest-related information dissemination (associated with Chapter 4), alter protest network and organisation (Chapter 5), as well as shaping the tactics and strategies of social movement and counter-movement campaigns (Chapter 6). However, only a limited amount of research has examined bot impact on the relations between protest and social media.

Therefore, the next sections of this chapter review existing literature on social media, protest, and social bots. The review explores the three dimensions of the impact of social media on social movements. In each section, it briefly reviews existing literature concerning the dynamic between protest and social media in this aspect. Then it summarises the few pieces of past work that explore the impact of social bots on the aforementioned dynamic. It is important to note two key considerations in this review. First, there are potential overlaps among the four subsections, but detailed discussion of each topic will be left to their respective sections. Second, similar to the dissertation itself, this section primarily focuses on protests within liberal democratic political systems. It will briefly refer to cases from other contexts such as authoritarian regimes but will not go into a detailed comparison of political and social media landscapes in different regimes.

### 2.2.1 Information dissemination

The widespread use of social media has fundamentally transformed how information about protests is disseminated, enhancing the speed, reach, and diversity of voices involved. In the past, traditional media outlets such as newspapers and broadcast networks controlled communication channels, playing a important role in facilitating the diffusion of protest-related information[109]. These outlets often filtered and framed the narratives surrounding protests, which influenced public perception and engagement.

However, the rise of the World Wide Web, particularly through social media platforms, has expanded the landscape of information dissemination. Unlike traditional media, social media allows a broader range of individuals, including ordinary citizens, to share their perspectives and experiences[5]. This democratisation of information flow has significantly impacted social movements, which relies heavily on effective communication and the rapid exchange of ideas.

Past research indicates that social media introduces two important changes to the dynamics of protest information diffusion. The first change is the reduction in the costs associated with sharing information. Social media enables instantaneous

communication, allowing protesters to broadcast developments in real-time. This immediacy not only keeps participants informed but also helps mobilise support quickly, as users can react to unfolding events without delays[110]. The second change is that social media platforms facilitate the rapid spread of information through mechanisms such as shares, retweets, and likes. A single post can quickly become viral and reach a large audience, amplifying the visibility of protest-related messages. This viral potential means that local protests can attract international attention, further increasing their impact and support[111].

Based on these changes, research at the intersection of protests and social media first further examines how information spreads through social media during contentious political events. Scholars build on established theories of opinion and information diffusion to investigate whether the dynamics of sharing information on social media—such as speed, reach, and audience engagement—differ in the context of protests and social movements compared to other forms of communication, like everyday social interactions or non-political events[109, 112, 113].

Two contrasting perspectives exist regarding the role of social media in information diffusion within protests and social movements. The first perspective claims that while social media facilitates protest-related information dissemination, the underlying mechanisms are not fundamentally different from those observed in traditional media contexts. Earl et al.[112] argues that three key factors drive information diffusion: how information is transmitted through media outlets, how content is framed, and the topics being covered. These principles apply equally to both social and traditional media. Similarly, agenda-setting theory suggests that media shape public opinion by influencing which issues are deemed important through selective coverage and emphasis[114]. Studies have shown that social media platforms are now sharing this agenda-setting power, previously dominated by traditional media[115, 116, 117]. Grassroots protest organisers can effectively 'set the agenda' of online discourse, attracting public attention and shaping sentiment around movements such as #BlackLivesMatter and #StopAsianHate[118, 117]. This, in turn, amplifies information diffusion and aids in protest mobilisation.

In contrast, the second perspective asserts that social media fundamentally transforms the dynamics of information diffusion in social movements due to its distinct characteristics. The design and functionality of social media platforms directly shape user behaviour and communication patterns during protests[119]. Unlike traditional media, social media is influenced by factors such as platform policies, user terms, community moderation, and the potential for anonymity, all of which create a unique environment for protest-related communication[120, 121]. These platform-specific features alter the mechanisms of information diffusion, either enabling or constraining how information spreads during protests, thereby influencing the effectiveness of online activism.

Furthermore, social media introduces diverse actors with distinct roles in protest communication, contributing to the spread of protest messages in unique ways[5]. Unlike traditional media, social platforms create intricate communication network structures, where the positioning of actors within these networks can significantly affect how widely and quickly information spreads[113]. For example, the ability to repost or share content can trigger rapid chain reactions, enabling protest messages to reach large audiences at an accelerated pace[122, 113]. In this environment, certain content, such as conspiracy theories and extreme views, tends to persist longer, which can have profound effects on protest communication as well[10].

In addition to the mechanisms of social media and how information spreads, a substantial body of prior research has focused on the outcomes of this information diffusion. Specifically, scholars have examined how it affects protest recruitment, mobilisation, and changes in the strategies and tactics used in protests. Much like research on the mechanisms of information diffusion, these studies often dissects one or more protests heavily utilising social media. One key finding is that online information diffusion is correlated with the geographical spread of offline protests across cities and national borders. From the earliest cases of social media facilitated protest, like Occupy Wall Street[123] and the Arab Spring[124], to later ones like Black Lives Matter[125], contentious communication on X and Facebook forecasted the eruption of offline protests.

Additionally, research has uncovered how online information diffusion alters the mechanisms of protest mobilisation. In the Indignados movement in Spain[126], repeated exposure to protest messages within users' personal networks played a important role in activating and motivating individuals to participate in protests. Similarly, during the Black Lives Matter protests[127], activists strategically used social media to shape and disseminate narratives around racial justice. This helped to educate the broader public and drive participation in offline protests by framing the movement' s goals and spreading calls to action.

Some studies have investigated the role of social bots in information diffusion on social media, particularly highlighting their activity and visibility within online discourse. Due to their highly automated nature and close network proximity, social bots generate a significant volume of messages in a short period and frequently retweet each other' s content to amplify their presence[128]. Additionally, they employ strategies aimed at maximising their reach, such as actively making connections with human users, as seen in certain cases[18]. This aggressive activity can draw public attention, influencing the flow of information online. For instance, Schuchard's research[94] revealed that a quarter of human users exposed to bot-generated messages during discussions related to the Ukraine conflict tended to retweet these messages. Further evidence from computer simulations supports the notion that social bots can successfully propagate information online[129], with empirical studies from the COVID-19 pandemic showing that the messages they spread often include rumours or fake news[70]. These studies have mapped the diffusion mechanisms using tools like agent-based modelling.

These research highlights important future directions, especially regarding how social media, social bots, protests, and information dissemination mechanisms interact. Although many studies have explored how information spreads during protests, the distinct characteristics of social bots suggest they might have a unique influence on this process. Social bots can operate at a scale and speed beyond human capabilities, potentially amplifying or distorting information flow in ways that

haven't been fully explored. Despite this potential, few studies have specifically investigated how social bots shape the spread of protest-related information.

Given this gap in the current literature, Chapter 4 of this dissertation will focus on addressing this issue by examining how social bots influence the flow of information during protests. This chapter will build on existing research while providing new insights into the specific role that social bots play in shaping communication and information flow online.

### 2.2.2 Network and organisation

Social networks represent natural forms of organisation that shape both information flow and connections between individuals, potentially impacting protests in the digital age. Furthermore, the emergence of new forms of online social networks (OSNs) on social media platforms has attracted substantial attention in the study of protests[122, 28]. Previous research on the intersection of protests and digital social networks has primarily focused on two key dimensions. The first dimension examines how network ties influence an individual's decision to participate in protests. The second dimension investigates the unique structural attributes of protest OSNs and their influence on protest dynamics.

In the first dimension, prior research investigated the influence of OSN ties on protest participation decisions at the individual level. This aspect builds upon a well-established body of literature concerning the impact of social network ties on political participation, a topic that was widely researched even before the widespread use of social media in social movements. In the era before social media's prevalence, studies primarily explored social ties in offline social movement participation and their impact on protest engagement[130, 83].

With the rise of social media, research in this area has shifted focus to two key areas. Initially, studies explored disparities in the characteristics of protesters within 'online' versus' offline' protest social networks. For example, Belgian online activists had stronger connections to formal organisations, while offline protesters relied more on informal ties[131]. However, findings from research on the Indignados

movement in Spain differed, as protests organised through online social networks lacked formal membership, with online-recruited participants being younger, more educated, and less politically engaged[126]. With the outburst of large-scale social media-facilitated protests, another research line investigated whether engagement in online social networks promotes individual protest participation offline. In general, individuals embedded in online social networks demonstrated a higher likelihood of participating in protests, particularly when actively involved in offline formal organisations[132, 133], as observed through survey results[134, 135], and in protest cases in France (such as the Charlie Hebdo protest)[136] and the #MeToo movement[137].

Next, an abundance of literature examined the structure of OSNs during protests and their impact on protest mobilisation and outcomes. In this context, 'online network structure' refers to OSNs formed by interactions on social media[122]. Research on protest networks typically employed social media users as network nodes, with edges representing retweets or following-followed relationships. Two primary methods were adopted to construct these networks. The first approach involved conducting keyword searches to identify all activities related to a specific protest event in the digital field and utilising the users involved in these activities as social network nodes[138]. The second approach began by identifying a group of active users and expanded the network scope for analysis based on their online social interactions[136].

Research on online social networks during protests predominantly focused on the characteristics and structure of these networks. They often identified common structural traits across various protest cases. At the individual level, protesters tended to be more interconnected compared to non-protesters, exhibiting direct, indirect, triadic, and reciprocated network ties[136]. Simultaneously, protest OSNs commonly displayed a core-peripheral network structure[124]. Highly active protesters occupy the central positions, and less active participants are located on the periphery. Despite the dense core, protest OSNs exhibited significant fragmentation at their periphery[139].

The unique network structure of protest online social networks (OSNs) significantly influences information exposure, protest organisation, and mobilisation, leading to effects that can either facilitate or impede information diffusion and mobilisation. The first effect examined is that the current OSN structure contains structural holes at its periphery, beneficial for information diffusion and protest activity spread[140, 124]. However, the fragmented nature of social media often hinders the full utilisation of these structural holes, creating echo chambers that impede information diffusion[122]. Furthermore, users at different positions of the network impact protest information diffusion and mobilisation differently. Gonzalez-Bailon et al. further identified 'broadcasters' and 'hidden influentials' in online protest diffusion[113]. Broadcasters are celebrities leading the diffusion process, while hidden influencers, central to information flow but at the margins of communication networks, hold distinct power in shaping the frame and movement identity.

Additionally, these structural traits not only facilitate online mobilisation but also potentially impact offline physical protest mobilisation and protest outcomes, as suggested by computer simulation[141] and a few pieces of empirical evidence[142, 136, 124]. However, the amount of research in this direction is relatively limited.

Existing literature on bots in OSNs primarily focuses on the network structure of social bot communities and the positions and behaviours of various agents. Most studies are based on case analyses of political events, including elections and international summits, with only a limited exploration of protest events. However, these studies shed light on bot behaviour and OSN structure within protest networks.

In terms of communities, research indicates that bots form dense communication networks ('botnets') by actively retweeting each other[80], which can be identified using community detection methods[77]. Botnets enable the rapid generation of large messages and information cascades with clear political preferences. Their messages often exhibit extreme political preferences and tones, as seen in the Brexit referendum[13] and the 2014 Japanese presidential election[97].

For the position and behaviour of agents in OSNs, studies offer differing conclusions about the position of social bots. For instance, some find bots centrally

positioned, making more attempts than human accounts to initiate contact through retweeting[143]. This activity characterises bots as active users in online political communication networks, such as in discussions related to Turkish government censorship on Twitter[94]. In scenarios such as the Gilets Jaunes (Yellow Vests) movement, bots appeared less central in social networks compared to verified accounts, as measured by network metrics[23].

There is only a limited number of studies for the other two aspects of protest network dynamics. There are no known studies examining bots' influence on offline protest participation. A few studies focused on bots' impact on human behaviour in social networks, mainly emphasising changes in public opinion. They suggest that bots attract more attention than verified accounts in protest networks[23]. Simulation and empirical data also suggest that a small number of bots in OSNs may effectively shift prevalent opinions in social media discourse[129]. However, existing social network theories also challenge this possibility by suggesting that users tend to be affected by their direct network instead of automated accounts in the periphery of their network[144].

These studies highlight future research directions for bots and social networks. First, there is a lack of studies examining bot behaviour in protest networks, essential for protest organisation and mobilisation. Second, existing studies primarily describe the position, attributes, and community of social bots in OSNs, with limited research on their impact on human-human edges and networks. Third, the few studies on bot impact in OSNs mostly focus on macro- or meso-level changes, which suggest a need for research on micro-level impacts, such as behavioural changes, network tie formation, and individual-level social network structural changes. Chapter 5 of this dissertation will address these issues accordingly, thereby addressing RQ3 (bot impact on human networks).

### 2.2.3 Mobilisation tactics

Social media has revolutionised how information about protests is disseminated and how online social networks are structured, leading activists to develop innovative

methods for mobilising and organising protests. A significant body of research has focused on how activists utilise social media to achieve their goals, with case studies highlighting influential movements such as the 2019 Hong Kong pro-democracy protests[145] and the #FridayForFuture climate strikes[146]. While each movement may use distinct strategies tailored to their specific context, common themes and techniques often emerge across various protests.

Caren et al.[5] identified four primary ways activists make use of social media. The first method is the mobilisation of offline protests or online actions through weak ties, as discussed in earlier sections. This section will explore the remaining three methods, supported by empirical evidence from various studies.

The second method involves sustaining communities on social media platforms for protest mobilisation and recruitment. Social media infrastructure allows activists to foster relationships, organise activities, and address negative responses from detractors[5]. By maintaining these online communities, activists can preserve collective identities that are important for cohesion among protest participants. Empirical studies in offline settings have highlighted the importance of community, identity, and cohesiveness in successful protest outcomes[147]. For example, during the Arab Spring, the online community on Twitter played a key role in mobilising Egyptians and supporting the Tahrir Square protests in 2011[148]. Similar patterns of community-building have been observed in movements like #MeToo[149] and far-right groups[150], which demonstrated how online platforms can sustain and extend communities across geographical boundaries.

The third method focuses on developing and spreading sympathetic movement frames. Building on framing theory, which originated in traditional media contexts, this approach involves creating and amplifying frames that resonate with the public. Benford and Snow[151] define frames as 'relatively coherent sets of action-oriented beliefs and meanings that legitimize and inspire social movement campaigns and activities.' Research in the social media context has shown that activists employ various tactics to craft and promote these frames, often in ways that differ from traditional offline methods[152]. For instance, in climate change protests, activists

have observed that framing processes on social media involve collective efforts where users highlight particular perspectives to influence public interpretation[153]. This process often results in more personalised frames that reflect individual experiences and emotions, rather than rigid collective schemes[154].

Finally, activists use social media to counter existing narratives about the issues they address. Hashtags on platforms like Twitter are a common tool for this purpose, as they help categorise conversations and frame discussions quickly. Hashtags can spread rapidly and facilitate viral information dissemination and mobilisation[127]. For example, the #MeToo movement used hashtags to challenge victim-blaming narratives in sexual harassment cases[155], while #BlackLivesMatter employed hashtags to protest police brutality[118]. Conversely, counter-movements such as #AllLivesMatter have emerged in response to #BlackLivesMatter, illustrating how hashtags can also be used to propagate opposing viewpoints[156].

Social bots, like human protesters, make use of a specific set of tactics to achieve their goals, which have been extensively reviewed in the subsection 'Behaviour and Strategies' above. These tactics often include communication strategies that are similar to those used by protesters. For instance, social bots frequently employ hashtags to either challenge prevailing narratives or promote particular messages. This tactic has been notably observed in discussions around controversial topics such as the BLM protest and the Brexit referendum[13, 157].

Despite the similarities in communication strategies between social bots and protesters, there is a notable gap in research regarding how bot strategies function within the context of protests. This lack of research leaves unanswered questions about the extent to which social bots' activities might intersect with or even undermine protest movements. Specifically, the impact of bot-driven narratives and activities on protesters' mobilisation efforts has not been thoroughly investigated. Given this gap, further research is needed to explore the dynamic interaction between the tactics employed by social bots and those used by protesters. Understanding how social bots' strategies might compete with, complement, or disrupt protest efforts is important for a comprehensive view of digital activism

and information dissemination. Chapter 6 of this dissertation will address these gaps accordingly.

## 2.3 Platforms

Both social bots and online activism rely on the infrastructure provided by social media platforms for their development and operation, which are subject to platform rules and other regulations. Consequently, platform design significantly impacts various aspects covered in this study, including the definition and tactics of social bots, the behaviour and strategies of activists on social media, the impact of social bots on protests, and the dynamic interactions between social bots and activists in the two online protest cases mentioned.

While platform design and platform rules are not the primary focus of this dissertation, understanding their influence is essential for a thorough analysis because they shape the context in which both bots and activists operate. Therefore, we will review how social media platforms impact social bots and how these platforms influence activism. This includes examining the mechanisms by which platform features affect bot behaviour and how platform rules impact both bot and activist activities.

Throughout this section, we will also justify our case study focus on X (Twitter), as it is frequently cited in the literature as a representative platform where bot activity and online activism are particularly prevalent. This emphasis on Twitter highlights its critical role in the dynamics of social media and illustrates the widespread recognition of its significance in understanding the behaviour and influence of social bots.

### 2.3.1 Platforms and bots

Social bots operate within social media platforms, and their existence is inherently tied to the design and infrastructure of these platforms. Technological advancements have enabled software to act as representatives of humans in various dimensions of online activity, particularly within social media. Users increasingly rely

on these platforms to shape their online identities and interactions, which social bots exploit to exert significant influence[158, 159].

The design of social media platforms allows bots to leverage specific features to appear more authentic. Past research highlights the concept of 'platform persuasion,' wherein social bots manipulate platform features such as follower counts, likes, and retweets to create an illusion of credibility and gain trust from users[49]. This tactic often involves using fake followers or engaging with vulnerable users, serving as a deliberate strategy to spread bot-generated messages more effectively[160]. Furthermore, bots can employ trending hashtags and participate in discussions to enhance their visibility, further blurring the lines between genuine user interactions and automated content[161]. Past work shows that misinformation disseminated by bots travels faster than truth-based content, which implies how platform design facilitates the rapid spread of bot messages[21].

Across the literature, researchers consistently discuss major platforms, with X (Twitter) frequently emerging as a focal point due to its unique features and high levels of bot activity. As mentioned, studies indicate that bots account for a significant proportion of accounts on Twitter[12]. This prevalence is particularly pronounced during major events, such as social movements, where bot activity can spike to amplify specific narratives or misinformation[23]. This consistent emphasis on Twitter not only highlights the platform' s critical role in the dynamics of social media but also illustrates the widespread recognition of its significance in understanding the behaviour and influence of social bots.

Although social media platforms are aware of the presence of social bots, their responses are shaped by competing motivations. While bots—including malicious ones—can increase user engagement and boost platform metrics, there is also mounting pressure for platforms to ensure authentic interactions. This tension is particularly pronounced during politically sensitive periods, such as U.S. elections, when platforms aim to avoid accusations of partisanship or facilitating misinformation[158, 162, 163].

Platforms like X (Twitter) also frequently emerge in discussions surrounding social bot rules, highlighting its critical role in this context. Stricklin et al. categorised platform rules into several areas, including automation, fake accounts, spam, and artificial amplification, all essential for distinguishing beneficial from harmful bot activities[158]. Specifcially, X (Twitter) has implemented policies to regulate bot activity while permitting legitimate automation. Notably, it provides APIs for developers to create bots and enforces user policies aimed at curbing malicious behaviour[164].

### 2.3.2 Platforms and activism

Social media platforms, particularly Twitter (now rebranded as X), play a important role in contemporary activism, much like they do for social bots, both of which rely heavily on platform infrastructure for organising, disseminating information, and mobilising people. Twitter has become especially important in activism due to its real-time content sharing, ease of use, and amplification capabilities[165]. This section reviews prior research on how social media platforms, particularly Twitter, influence digital activism.

The concept of 'platform affordance' is commonly used in research to explain how platform infrastructure affects activism. Although platform affordance applies to various social media platforms, a large number of research in this field focuses on X (Twitter) due to its distinct design features and widespread use in political and social movements. Platform affordance refers to the perceived or actual capabilities a platform provides to users, shaping what they can and cannot do[166].

In Twitter' s case, affordances such as hashtags, retweets, trending topics, and the original 140-character limit have proven critical for digital activism, enabling the rapid spread of information and broad participation[165]. Therefore, past work examined specific Twitter features and how they shape activism. Seigert[167] examined how the four affordances that distinguish social media from traditional media-visibility, persistence, association and editability-demonstrated in the case of Twitter. Jaidka et al.[168] explored how Twitter' s original 140-character limit

influenced political discourse by encouraging brevity, which shaped how activists framed their messages. This characteristic of Twitter forced activists to communicate more concisely, making their messages more digestible and easier to share widely. Similarly, Rathnayake et al.[169] highlighted how Twitter' s hashtags features enable activists to amplify their voices and build momentum for their movements quickly. While platform affordances apply across social media, Twitter' s unique design makes it a focal point in these case-specific studies.

Beyond examining platform design, case-specific studies have investigated how platform affordances shaped activism during particular events. Because of the prevalence of X (Twitter) for activists, most of those studies analysed the role of it in shaping activism cases. Earlier examples include studies on the Arab Spring[120] and the Gulf of Mexico oil spill protests[170], while more recent research has investigated queer activism[171] and the Iranian women's rights movement[172]. These studies highlight how specific affordances—like ease of information dissemination and organisational capabilities—affect identity formation, information flow, and movement outcomes in unique protest cases.

In summary, past work on platform and bots, as well as platform and activism, extensively highlights the importance of X (Twitter) to both bots and activism. Twitter' s critical role in activism is well-documented, with its unique affordances. Similarly, the platform serves as a fertile ground for social bots, which can exploit its features to engage users, spread misinformation, and amplify specific narratives. Given this dual significance, studying the interplay between bots and activism on Twitter is particularly important. However, there is a notable gap regarding the interaction between bots and activism on this platform. Thus, this study chooses to focus on bot activity on X (Twitter) and in online activism to explore the relationship between bots and activism.

## 2.4 Context of protest cases

This research focuses on two major ongoing protest incidents for case studies: Extinction Rebellion (XR) and Black Lives Matter (BLM). As detailed in Chapter 1, these cases are representative of contentious political movements that extensively make use of social media, with significant traces of automation. Both environmental and racial issues are prominent topics in elections, news media, and civic activism, making them important for understanding contemporary activism.

Examining the role of automation in these two protest cases, as covered in Chapters 4, 5 and 6, is vital for comprehending the dynamics of modern activism related to environmental and racial concerns. To effectively analyse these cases, it is essential to provide a thorough overview of their backgrounds and past research. This background information will help contextualise the impact of automation in these cases and highlight the significance of these movements in the broader landscape of social and political activism.

### 2.4.1 Extinction Rebellion (XR)

Extinction Rebellion (XR) is a global grassroots environmental movement that emerged in the UK in 2018 and has remained active since then[44]. Originating from a call to action by scholars in May 2018, the movement gained momentum with a sit-in at the Greenpeace UK headquarters on 17 October, 2018. This marked the beginning of widespread protests across the UK and globally, particularly active from 2019 to 2020 in the UK, US, and Europe[173].

XR identifies as a grassroots, non-partisan, participatory, decentralised, and inclusive social movement[44]. Activists organise autonomously in small groups globally, adhering to XR's goals and principles. Critics argue XR lacks adequate representation of people of colour and low-income individuals significantly affected by climate change[174]. The movement is organised into 'anchoring circles,' small teams overseeing various aspects of activism, such as actions, logistics, fundraising, lobbying, and advocacy[44].

Funding for XR comes from crowdfunding, major donors, NGOs, and foundations. The majority of 2019 funding came from individual donors through crowdfunding, supplemented by contributions from the Climate Emergency Fund, the Children's Investment Fund Foundation, and various US philanthropists[173]. Funding distribution is controlled by anchor circles unless delegated otherwise.

XR's primary goal is to use non-violent disobedience tactics to highlight ecological destruction and climate change, urging governments to take action[44]. The UK movement demands that their government address climate change, halt biodiversity loss and greenhouse gas emissions by 2025, and create a Citizens' Assembly for climate and ecological justice decisions[175].

To achieve these goals, XR employs various tactics, including occupying public spaces, causing intentional disruption, lobbying Parliament, and willingly facing arrest. Since winter 2019, XR has organised non-violent disruptive events across the UK and globally, primarily targeting capital cities[176]. Social media platforms such as Facebook, Instagram, and Twitter play a important role in XR's communication strategy, helping them reach the public and disseminate their message[45].

To understand the context and significance of Extinction Rebellion (XR), past work first examined the historical and contemporary discourse surrounding climate change activism. Prior research and media sources[177, 9] have highlighted XR as a contemporary social movement deeply rooted in the longstanding history of environmental and climate change protests. Since the 1970s, climate change has been recognised as a critical social issue, leading to major protests, notably in the 1990s[178]. Some scholars argue that the 2009 United Nations Framework Convention on Climate Change (COP15) summit in Copenhagen marked a pivotal moment for global climate activism, sparking widespread civil engagement[179, 180]. The subsequent COP21 summit in Paris further intensified mobilisation and drew attention to the need for pressure on international climate negotiations[181]. However, as noted by DeMoor et al.[178], climate change movements have increasingly shifted their focus to challenge the perceived inadequacies of these negotiations. Activists

now strive to build a movement capable of addressing climate issues independently, beyond the scope of summit outcomes[182].

These aspirations and objectives gave rise to waves of climate change activism, such as movements like Fridays for Future (FFF) and XR. FFF originated when the 15-year-old high school student Greta Thunberg skipped school to protest climate change in front of the Swedish Parliament, urging the government to adhere to the Paris Agreement and take stronger action against global warming. Thunberg's activism inspired students and activists globally to join the movement. Similar to XR, scholars argue that these movements are distinct from past protests against environmental issues, including those surrounding the COP[181].

Consequently, existing studies on Extinction Rebellion (XR) have largely concentrated on understanding what differentiates XR from previous climate change activism. These studies have focused on two main aspects that set XR apart: its unique discourse, narrative, and framing, and its innovative organisational strategies.

The first area of research has focused on how XR frames its movement, examining the discourse and narrative constructed by activists related to their goals and activities. Scholars draw from waves of environmental and climate change activism to compare XR with previous protests, emphasising the diagnostic framing shared with other climate change movements[183]. XR, like its predecessors, identifies ecological destruction as a pressing issue, employing narratives of 'vernacular eschatology' and 'postapocalyptic' to mobilise the public towards a better future[184]. XR activists have also crafted a distinct narrative by integrating established climate change framing with elements from other political ideologies[185]. For instance, XR has redefined the concept of 'radical protest,' challenging traditional academic definitions that equate radicalism with violence. They argue that their peaceful protests are radical in nature, reinterpreting radicalism as a stance that challenges conventional norms[186]. Additionally, XR's framing emphasises climate change justice and places a stronger focus on state actors rather than non-state entities.

The incorporation of vanguardism, a concept often associated with leftist politics, underscores XR's efforts to adapt and 'green' its environmental agenda[187].

The second category of research focuses on XR's mobilisation repertoire and organisational strategies. Scholars researched the tactics employed by XR activists, such as non-violent disobedience, aimed at raising public awareness[188]. However, past research argued that the scale of protests was more likely to have attracted public attention[178] compared to its tactics. Scholars also highlight inconsistencies between XR's movement framing, which emphasises systemic change, and the lack of specific goals in climate change protests, potentially undermining the systemic change they advocate[189].

While previous research has acknowledged the broader influence of social bots on online climate change activism, there is a clear gap in studies specifically focusing on their role within Extinction Rebellion (XR) discourse and activism. Most existing research on XR has concentrated on the movement ' s framing and mobilisation strategies, examining how activists develop their narratives and influence public attitudes towards climate change. However, with the growing presence of social bots on social media platforms, it is important to investigate how these automated entities interact with and impact XR activists' discourse.

Social bots have the potential to significantly sway public opinion and alter the dynamics of online activism, which can affect how XR is perceived and engaged with by the public[70]. Therefore, a detailed exploration of social bots' impact on XR is necessary. This includes examining how social bots influence public support for XR (discussed in Chapter 4, which also serves as the foundation for addressing RQ1(bot impact on human behaviour)). Furthermore, Chapter 6 will examine the strategies employed by these bots to achieve such impact, specifically in the context of XR, which also addresses RQ3(bot strategies)). By addressing these aspects, the study will bridge the gap in current research and provide a comprehensive understanding of the interplay between social bots and XR ' s online presence.

### 2.4.2 Black Lives Matter (BLM)

The Black Lives Matter (BLM) movement traces its origins to the #BlackLives-Matter hashtag, which emerged in July 2013 following the acquittal of the white police officer George Zimmerman in Trayvon Martin's death. Street demonstrations began in 2014 in Ferguson, Missouri, after the death of Michael Brown, marking the movement's transition from social media campaigns to widespread activism.[190] BLM primarily addresses issues related to government violence against Black Americans, especially police killings. Activists aim to combat racism, discrimination, and racial inequality, shedding light on unfair treatment and structural inequalities. The movement advocates for improved police training, accountability for misconduct, and the reallocation of police funds to community services[191].

Since 2014, the BLM movement has gained prominence, evolving into one of the largest ongoing social movements in the United States. Racial unrest events, often triggered by incidents of police violence against Black Americans, occur annually. The movement has expanded globally to countries like Australia, Brazil, Canada, and various European nations. BLM activists asserted that these activities are driven by a decentralised group of individuals, emphasising the absence of a formal hierarchy in their organisational structure. It is worth noting, however, that several organisations use the label 'Black Lives Matter,' such as The Black Lives Matter Global Network Foundation.[190, 192]

The concept of BLM originated on social media, and BLM activists heavily relied on platforms like X (Twitter) as a important means to disseminate their messages and call for actions both nationally and internationally[138, 193, 127]. In addition to making use of social media, they combated racism through various channels, including political actions, petitions, campaigns, and protests. Offline collective actions primarily involved mass protests, demonstrations, and other contentious events[194].

As one of the major social movements in contemporary United States, the BLM movement has generated significant controversy. The movement has attracted a diverse array of supporters, and has also provoked substantial criticism directed at

its goals, slogan, and organisation[195]. Concerns have been raised about the 'Black Lives Matter' slogan, with some feeling it might overshadow the importance of lives beyond the Black community. This concern led to the development of the counter hashtag #AllLivesMatter and a series of counter-protests[196]. Additionally, there have been reports of violence during several major BLM protests, where non-violent demonstrations during the day are followed by nighttime incidents of riots, arson, and robbery[197]. This has contributed to a perception among a portion of the general public and news media that BLM is a violent group targeting law enforcement and white individuals[198].

Moreover, BLM has had a significant impact on various political processes. It became a focal point of debate during the 2016 and 2020 presidential elections, with different political parties expressing polarised opinions on the movement[199]. Beyond elections, the movement has influenced other political processes, including legislation and policymaking, such as the cutting down of political budgets[200].

This dissertation project specifically focuses on the eruption of protest events that occurred in late May to early June 2020 following George Floyd's death on May 25, 2020, in Minneapolis, Minnesota. George Floyd, an African American man, died while in police custody when an officer knelt on his neck for nearly nine minutes during an arrest. The incident, captured on video and widely shared on social media, sparked widespread outrage, calls for justice, and led to extensive protests and demonstrations across the United States[46].

BLM is extensively studied in online activism due to its historical significance, widespread impact, and extensive use of social media. Research on BLM primarily falls into three categories: its historical and cultural roots, organisational formation and activities, and the role of social media in protest mobilisation.

The first category of research on BLM investigates into its historical and cultural origins. Researchers explore how past events, notably racial violence dating back to the Reconstruction amendments, contribute to framing the contemporary movement. While BLM focuses on addressing political killings nowadays, its roots trace back to racial violence and structural inequalities faced by African Americans

since the Civil War[201]. For example, Lebron[202] investigated how African American intellectual, cultural, and artistic traditions shape the thought history behind BLM. They then emphasised how the distinct thought origins of BLM significantly influence its framing and narrative. As previously mentioned, BLM emerged to confront ongoing racial violence, with the goal of altering both public discourse and policies surrounding these issues[201]. Extensive studies reveal that BLM activists predominantly present the movement as a quest for individual rights. Therefore, they incorporated existing frames related to gender, race, and LGBTQ identities[190]. BLM strategically portrays itself as a tribute to historical Black Liberation struggles, redefining the narrative to encompass all Black lives, recognising differences in gender, sexual orientation, age, nationality, or criminal status[203].

The second category of past literature centres on the organisation, network, and mobilisation of BLM. Previous research has demonstrated that both offline and online social connections play a important role in mobilising individuals for participation in the BLM movement[138, 204, 205]. Due to the focus of this dissertation, the discussion will primarily summarise research on the digital network of BLM activism on social media, which has three major focuses.

First, BLM network studies involve examining the network structure. Bennett et al.[206] argued that, similar to other movements on social media, the online communication network of BLM exhibits a dense centre with many loosely connected components. Second, some past work focuses on categorising content and sentiment of discussions using social network methods. Studies in this aspect employ community detection methods to identify different focuses of discussion in protest-related discourse, such as black empowerment discourse, black culture advocates, and liberal activists[207]. Previous work in this aspect has also revealed digitally networked extreme right opponents that synchronise online, with discussions on their network structure and content analysed using network analysis methods[208]. Third, research explores the impact of such network structures. It is noted that such network structures effectively enable hashtags to spread through networks, facilitating information diffusion[118]. However, computer simulations

indicate that these networks may promote participation in high-cost behaviours while hindering participation in low-cost behaviours[204].

The final dimension of past work on BLM deals with its relations with social media, as the movement has been very prominent on social media since its very beginning. Past work in this perspective extensively studied the tactics of how BLM activists make use of social media. Most studies acknowledge that BLM activists and supporters extensively engage in the so-called 'hashtag activism.' Therefore, a large proportion of studies on BLM social media tactics focused on how BLM activists used hashtags (#BlackLivesMatter, #Ferguson, etc.) to mobilise the BLM movement. They utilised those hashtags to facilitate solidarity, refer to police violence, discuss movement tactics, disseminate information about specific police killing events such as Ferguson and Floyd' s death[127]. At the same time, counter-movement sentiment was also expressed through hashtags[156]. Those hashtags seemed to have diffused widely and had a significant and stable effect on shaping the public agenda after two weeks, and such effect remained stable afterwards[118]. A few pieces of past work also examined other factors that contribute to the online mobilisation of BLM, such as the involvement of (black) celebrities[209, 210].

Reports from various media outlets and a limited number of existing studies have discussed the resurgence of bots during BLM protests. Past research indicates that bots were notably active under the #BlackLivesMatter hashtag, particularly following the death of George Floyd[211]. Different social bots played different roles in the protest related discourse. Some expressed support for BLM and criticising police actions, while others engaged in activities aimed at spreading controversy and disinformation[212]. Notably, some social bots associated with the Internet Research Agency (IRA), an entity accused of conducting information operations in support of Russian political interests, were identified participating in BLM communication.

While existing studies have explored the identification and behaviour of social bots during the BLM protests[211], a significant research gap remains in under-

standing how these bots impact the structure and dynamics of human online social networks.

This gap is particularly notable because past studies have extensively analysed human online social networks in the context of BLM, focusing on how individuals and groups interact and mobilise[208, 204]. However, there has been little investigation into how the presence and activities of bots, which often form complex networks of their own[13], impact these human social networks. Understanding this interaction is important because it can reveal how bots might alter or disrupt the dynamics of online activism and influence public engagement and perception.

Chapter 5 of this dissertation aims to fill this critical gap by examining how social bots affect the structure and dynamics of human online social networks during the BLM protests. Filling this gap serves as the foundation for addressing RQ2 (bot impact on network). By exploring the interplay between automated entities and human users, this chapter will provide valuable insights into the broader impact of bots on digital activism and the ways in which they shape and influence human social interactions and network dynamics.

Furthermore, Chapter 6 will expand on this by providing a detailed overview of bot activities during the George Floyd protests, demonstrating what kind of bot activities impact human communication and further contributing to the understanding of their influence on social networks.

In summary, this chapter has provided a comprehensive literature review that addresses the central research question: How do social bots impact online activism? This review has examined the key actors, including social bots, activists, and social media platforms, and how they come together in the two protest cases we examined. By doing so, it has established a foundational understanding of their interactions and the dynamics that shape the landscape of digital activism, and has identified research gaps accordingly.

With this comprehensive overview of existing research, the identified gaps, and contextual examples, this study will then embark on the analysis of social bots and their impact on online activism. The following chapters will explore these dynamics in greater detail, providing empirical insights and theoretical advancements that contribute to our understanding of this critical intersection of technology and civic engagement.

# 3 Methodology

## Contents


**3.1 Data collection . . . . . 59**
**3.2 Bot identification . . . . . 62**
3.2.1 Botometer as the benchmark Model . . . . . 63
3.2.2 Self-trained models . . . . . 66
**3.3 Text categorisation using ChatGPT . . . . . 68**
**3.4 Topic modelling . . . . . 69**
**3.5 Approach to identifying potential causal relationships . 72**
**3.6 Conclusion . . . . . 75**


This chapter introduces the methodology that serves as the basis for all analyses conducted in the three empirical chapters of this dissertation. It encompasses five main aspects of the methodology: data collection, bot identification, text categorisation using ChatGPT, topic modelling, and approach to identify potential causal relationships. The first part explains the procedures used to gather raw data, including tweets and user profiles. The second part covers the approach used for bot identification, which forms the foundation of the empirical analysis. The third part describes text categorisation with ChatGPT, which this dissertation uses to classify opinion of protests. The fourth part, topic modelling techniques for textual data analysis, involve data preprocessing and bi-term topic modelling. The fifth and final part presents the research design that identifies association and potential

causal relations between bot activity and human behaviour change. This framework is generic and has been applied to both Chapter 4 and Chapter 5. Within each of these chapters, specific research designs have been tailored to align with the unique context and objectives of each study. This chapter discusses the validity of each method and details the robustness checks conducted on the methods, to ensure their effectiveness, replicability, and reliability. The final section (conclusion) outlines the main methodological contributions made through this dissertation project, from initial data aggregation to sophisticated bot detection, natural language processing, text categorisation with large language models (LLM), and causal inference designs.

**Author contributions:** See Appendix D for signed author contribution declarations.

## 3.1 Data collection

The research centres on the examination of specific protest cases and their associated political communication dynamics. To ensure a comprehensive analysis, this research used a complete archive of tweet data, which provides a holistic view of the entire life-cycle of these protest instances, including any bot-related activities and their potential influence on contentious communication. The selection of these cases was guided by a combination of epistemological considerations and practical constraints. These factors encompass the inherent characteristics of the protests, the individuals or groups representing them, and the feasibility of conducting empirical research.

To achieve this, the dataset consists of Twitter activity related to two periods during the protest events of interest. The first event is made up of multiple waves of Extinction Rebellion (XR) climate change protests that occurred between 18 November 2019 and 10 December 2019. XR is a worldwide environmental social movement that originated in the United Kingdom in October 2018[44], and has remained active ever since. This research centres on the analysis of social media discourse pertaining to multiple waves of XR climate change protests that occurred from late October to early December 2019. The second event involves the outburst

of Black Lives Matter (BLM) protests following George Floyd's death on 27 May 2020. Black Lives Matter (BLM) is an activist movement that commenced in 2014 in Ferguson, Missouri, in response to street demonstrations following the tragic death of Michael Brown on 9 August, 2014. Its inception was driven by the need to raise awareness about police killings of unarmed African American citizens. Since its inception, the movement has persistently worked to combat racism, discrimination, and racial inequality. The passing of George Floyd in Minneapolis, Minnesota on 25 May, 2020, led to a notable resurgence of the Black Lives Matter (BLM) movement. Tweet data posted during the peak of the protests over a 72-hour period was collected.

For the XR protests, data collection was completed in November and December 2021. Data for the BLM protests were obtained from August to September 2022. The data collection process made use of the Academic Research product track application programming interface (API) provided by Twitter, which was available at the time of data gathering[213]. This API allowed for the collection of a full sample of historical tweets based on specific keywords and conditions. To gather relevant data for both protest cases, this study collected all tweets posted during the specified time periods containing keywords related to the protests. For example, for the XR case, keywords such as 'Extinction Rebellion,' 'climate change protest,' 'XRebellion,' 'XR,' and various similar variations of these keywords were used. The complete list of keywords used can be found in Table 3.1. Initially, these keywords were selected by observing the trends of protest-related hashtags. After the first batch of 2,000 tweets was collected, the keyword list was further refined by adding protest-related keywords and hashtags identified from the sample. Alongside the tweets, user profiles of those who posted the tweets were also collected. Multiple features relevant to the subsequent analysis were queried.

The data collection method was determined by drawing upon established methodologies from prior studies[113, 25, 28] as benchmarks. Alternative techniques for identifying Twitter activities exist. These techniques initially sampled active users during protests, employed snowball sampling to gather profile details of all

activists, and subsequently collected all the tweets within their timelines[136, 124]. A similar approach was adopted for subsequent data collection, tailored to the specific research designs outlined in Chapters 4 and 5. However, initial data gathering still begins with a keyword-centric approach, prompted by the following rationale. This research examines bots and their influence on online discourse and information dissemination. Given this focus, it is reasonable to commence with sampling an exhaustive archive of discussions, rather than starting with users and their interconnectedness.

The final database consists of two datasets. The XR dataset contains 201,010 tweets and 122,130 users, while the BLM dataset contains approximately 4,000,000 tweets and around 2,000,000 users. Further data collection was conducted as part of specific research designs, which will be elaborated on in Chapter 4, 5, and 6 accordingly.

As the API provided access to a full archive of tweets, this is regarded as a fairly comprehensive sample of all political discourses related to the protest discourse of interest on Twitter. However, it is important to note that the data was collected retrospectively. Twitter applies some level of censorship on spam messages, and users might have deleted their accounts or messages between the protest events and the time of data collection. Due to the controversial nature of the subject (protests), the volatile behaviour of social bots, and the fact that bots are often disposable (tending to be used only once), there is a higher possibility for social bot messages to be deleted or accounts to be suspended compared to human users. For example, an active bot account run by XR activists was suspended at the time of data collection, making it impossible to retrieve the user profile or timeline. Consequently, the estimation of the scale and impact of bot activities might be lower than reality.

Finally, it is important to mention that the Twitter Academic Product Track API has undergone major changes in the past two years. It was initially unavailable, and when reintroduced, it came with a paid access model[214]. The data for this dissertation was collected before these changes took effect. Nonetheless, these

**Table 3.1:** Keywords collected for data collection.

| Case | Keywords |
|---|---|
| **Extinction Rebellion** | Extinction Rebellion |
| | #ExtinctionRebellion |
| | XRebellion |
| | XR |
| | climate change protest |
| **Black Lives Matter** | BLM |
| | BlackLivesMatter |
| | Black lives matter |
| | George Floyds |

Note: All keywords are case-insensitive.

recent developments highlight the growing challenges in accessing social media data, which makes the timing of this research particularly valuable.

## 3.2 Bot identification

This dissertation project centres on the examination of social bots and their influence on individuals within the realm of contentious political communication. The methodological foundation of the research design is bot identification. A significant methodological advancement within this dissertation project is the development of a viable, robust, and scalable approach for identifying automated agents on social media platforms.

To identify political bots on Twitter, this study used a combination of two sets of bot identification methods. The first is a state-of-the-art Twitter bot identification tool known as Botometer (formerly BotOrNot)[35, 215], which was used as a baseline and benchmark. However, the Botometer tool faced criticism for its false positive problem[36]. Furthermore, it is a generic tool used for all types of social bots, while the research focuses on social bots in political communication, especially activism. Even though Botometer was widely used in the political context[72, 216], past works still argue for the necessity of political-context-specific bot detection methods[72, 23]. Consequently, this study proceeded to implement, fine-tune, and compare several alternative algorithms for detecting social bots, in particular using datasets of political bots as training sets.

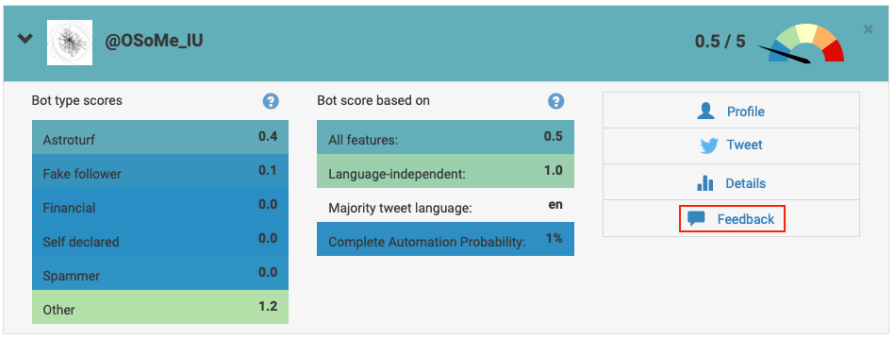



**Figure 3.1:** Sample output of Botometer bot classifications outcomes.

To construct the final bot detection pipeline, this study combined the two effective bot detection methods. In this section, the approach adopted to refine both the outcomes from Botometer and the self-trained classification algorithms is detailed. Subsequently, it outlines how these two techniques were integrated to establish a robust, replicable, and scalable method to detect social bots on Twitter, especially in the political communication context.

### 3.2.1 Botometer as the benchmark Model

The bot identification process began by adopting the widely used Twitter bot identification tool known as Botometer. This publicly available tool relies on machine learning and was initially released on 1 May 2014, with the latest update in September 2020. The developers of Botometer explain that it is designed to calculate a score where low scores indicate likely human accounts, while high scores suggest likely bot accounts[35]. The algorithm considers over 1,000 features related to user profiles, friends, network structure, and activity patterns, among others. Due to its widespread use by social scientists for identifying Twitter bots in research[72, 216], Botometer was adopted as the baseline model for bot identification. The results were mostly used to compare with the self-trained models.

The Botometer tool works as follows. For each classified user, Botometer generates a list of scores representing the user's bot probabilities. The scores include an overall score and sub-scores categorised based on the nature of bots: echo-chamber, fake follower, financial bots, self-declared bots, spammers, and others.

An example of the bot scores provided by Botometer for a given user is illustrated in Figure 3.1. According to the developers, these scores, also referred to as complete automation probabilities (CAP), represent the probability that an account with a score equal to or greater than the given value is controlled by software[35]. For instance, a user with an overall CAP of 0.96 means that 96% of users with similar profiles as theirs are fully automated. Since the research design required binary bot identification results, this study followed the common practice observed in existing studies[72, 215, 24, 14]. A threshold was defined, and users with an overall CAP exceeding this threshold were considered 'bots,' while those below it were classified as 'humans.' This approach allowed the distinction between automated and human-operated accounts.

However, considerable debate surrounds the appropriate usage and interpretation of these scores[36, 72]. To address the controversies and concerns, a series of robustness checks were conducted. The primary concerns are about determining the best practices in interpreting the Botometer scores. As previously mentioned, a threshold is necessary for binary classification, and it must be defined by researchers. Existing studies have employed various standards, ranging from 0.25 to 0.76. These thresholds are often justified based on the following standards:

a) Common practices in machine learning, where 0.5 serves as a widely used threshold for binary classification tasks.

b) A 'tipping point' approach, where the threshold is set at a value where the majority of users' bot probabilities fall below it.

c) Insights derived from the Botometer team's reports on model performance on their training data, utilising metrics such as the F1 score and Receiver Operating Characteristic - Area Under the Curve (ROC-AUC).

d) Tailoring the threshold to the specific context and research needs of the study.

By exploring and accounting for these different threshold options, this study aimed to ensure the robustness and accuracy of the bot identification methodology. This approach allowed the classification to adapt to the unique context of the research, enhancing the reliability of the findings.

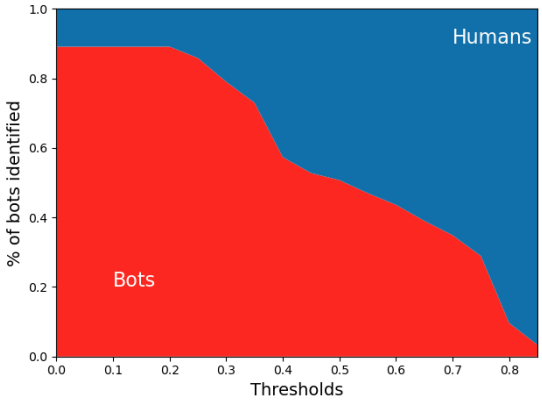



**Figure 3.2:** Proportion of users classified as 'bots' (red) and 'humans' (blue) in the dataset, with bot thresholds ranging from 0 to 1.0.

Based on the methods adopted by previous studies, this study chose to use the thresholds of 0.65 and 0.5 to classify bots. Specifically, users with a CAP higher than 0.65 were categorised as 'automated,' those with a CAP less than 0.5 as 'human,' and those falling in between as 'unknown.' This decision is based on the following rationales: First, 0.5 was adopted as the threshold for humans, following common practices (method a above). Second, the research focuses on investigating the impact of social bots on human behaviour at both the individual (micro) and network (meso) levels. Therefore, the validity of the results is particularly sensitive to false positives (incorrectly classifying humans as bots). As a result, this study selected a more conservative threshold for bot identification to avoid false positives. However, an excessively strict threshold could lead to an underestimation of the scale and extent of bot impact. To address this concern, this study adopted an improved approach described in standard b) above: bots were classified using the strictest threshold possible before the tipping point that was worked out, ensuring that a threshold leading to an excessive amount of false negative classification outcomes was not used. Figure 3.2 illustrates the distribution of users classified as 'bot' or 'human' at different thresholds in the case of XR Twitter discussions. After considering the aforementioned rationales, 0.65 was determined to be the most suitable threshold for bot identification in this study.

**Table 3.2:** List of the training datasets.

| Dataset |
|---|
| Twibot-20 |
| verified-2019 |
| vendor-purchased-2019 |
| political-bots-2019 |
| cresci-rtbust-2019 |
| botometer-feedback-2019 |
| celebrity-2019 |

### 3.2.2 Self-trained models

To address the concerns associated with solely relying on Botometer for identifying bots (which can result in false positives and lack contextual specificity), the study also incorporated supervised machine learning models. These models were trained using publicly available data that encompassed both bot accounts and human users. The training sets were sourced from existing open-source data of social bots in political communications. Table 3.2 provides a list of all the training datasets that were used. These datasets encompassed various types of political bots identified through previous research and user feedback, such as fake followers, spam message bots, and astroturfing bots. Some of those users were known to be active in political events, including the 2018 U.S. midterm election. The data was divided into a 70% training set and a 30% testing set. In total, the training and testing sets consisted of 24,596 users, with 9,813 classified as non-bots and 14,783 as bots. Since only 40% of the training set comprised bots, appropriate weights were applied to the training dataset for both bots and non-bots to ensure balanced learning. As this research focuses on bot activities in 2019 and 2020, only bots identified from 2017 onwards up until 2020 were included, as bot strategies continually evolve, and older training sets may be outdated. Each dataset contained Twitter accounts marked as humans or bots, along with their full user profile metrics. To train all models effectively, grid search methods were employed to determine the optimal parameters for each model, ensuring robust performance and accurate identification of political bots.

To obtain the best-performing model, several machine learning models were evaluated. These models were selected based on their performance in similar classification tasks as documented in previous works. The models evaluated included Decision Trees, Random Forests, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and neural networks. The models were trained with the aforementioned datasets, and their performances were compared using the ROC-AUC metric. The results indicated that the GBM and Random Forest models outperformed others with ROC-AUC scores exceeding 0.85. Due to their strong performance, these two models were selected for further analysis.

To verify the classification outcomes of the GBM and Random Forest models, this study also performed an in-depth audit of the models using SHAP (SHapley Additive exPlanations) values[217]. The SHAP values allowed for the decomposition of the output of each model into individual feature contributions, offering a detailed understanding of the decision-making process of the models. This analysis ensured transparency in the model's classification outcomes and confirmed that the models effectively identified features commonly associated with bots in political communication.

After identifying the best-performing models and validating their results, this study combined the results from the Botometer benchmark model and the self-trained GBM and Random Forest models to enhance the accuracy of the bot identification process. Users who were classified as bots by both Botometer and the self-trained models were labelled as 'bots' with high confidence. In contrast, users flagged as bots by only one method were subject to further scrutiny and manual inspection before being categorised. This ensemble approach, combining both Botometer and self-trained models, provided a robust and scalable method for identifying social bots in political communication contexts on Twitter, minimising the risks of false positives and negatives.

## 3.3 Text categorisation using ChatGPT

This study required a series of NLP classification tasks in the empirical chapters focusing on social media posts from both bots and humans. These tasks included categorising users' support status on protests (Chapters 4 and 5) and classifying bot activity based on the framework devised in this research (Chapter 6). This study employed OpenAI' s large language model (LLM), ChatGPT 3.5[218], to complete these tasks. While specific procedures for textual classification are discussed in the respective chapters, this section provides an overview of the method selection and its validation.

Text classification using Large Language Models (LLMs) has gained popularity in recent research, with various studies adopting these models for their ability to handle complex textual data[219, 220].LLMs have been especially effective in analysing social media and political communication, where understanding complex contexts and background knowledge is often essential. Traditional NLP methods often struggle in these areas due to their limited ability to interpret the context-dependent languages used in social media and political discourse. In contrast, LLMs can capture these subtleties, making them more effective for tasks that require deep contextual understanding[221, 222]. These models have demonstrated strong performance in tasks such as sentiment analysis[223, 224] and topic modelling[225].

There has been extensive discussion regarding the validity and performance of LLMs in these contexts. Research suggests that LLMs can generalise well across different datasets and domains, often outperforming traditional machine learning methods that require significant feature engineering[226, 222]. Given their success in empirical studies and ongoing improvements, LLM-based text classification is considered a robust and effective method for analysing complex communication patterns on social media[227, 224].

To measure the validity of ChatGPT-based classification in the specific context of this research, this study took a random sample of $N = 300$ users and asked two independent coders to replicate ChatGPT' s classification in one sample task:

classification of users' standpoints towards BLM and XR (results used in Chapters 4 and 5). The information provided to the coders was identical to the input data given to ChatGPT—combined timelines of users. Table 3.3 shows the correlation between ChatGPT output, the average score of the two coders, and the correlation between the coders and with ChatGPT separately. The results of the two coders showed a high correlation ($C = 0.71$), and their averaged opinion correlated even higher with ChatGPT's answer ($C = 0.88$). Therefore, it was concluded that ChatGPT's support classification is a valid method for automatically inferring users' support levels.

**Table 3.3:** Pearson Correlation between the support level values of Twitter users labelled by ChatGPT and manual coders.

| | ChatGPT | Average score | Coder 1 | Coder 2 |
|---|---|---|---|---|
| ChatGPT | 1.00 | 0.88 | 0.72 | 0.91 |
| Average score | 0.88 | 1.00 | 0.92 | 0.93 |
| Coder 1 | 0.72 | 0.92 | 1.00 | 0.71 |
| Coder 2 | 0.91 | 0.93 | 0.71 | 1.00 |

The procedure for ChatGPT-based classification was as follows. For each user, this research provided the model with an instruction prompt detailing how to perform the classification tasks. The LLM was also provided with the text to classify, usually the users' tweets. The model then outputted the aforementioned classification results. For specific tasks, data, and results, please refer to the methodology section and appendices for Chapters 4, 5, and 6 accordingly.

## 3.4 Topic modelling

The dissertation is embedded within the context of online political communication. Conducting research on this topic requires a comprehensive understanding of protest-related discourse content, opinions, and trends in both XR and BLM cases. To achieve this understanding, it is important to first identify the primary topics and communication agendas surrounding these protests. This involves analysing the discourse and information flow from both bots and humans separately. Such

analyses not only shed light on bot activities but also provide valuable contextual insights that inform subsequent bot activities. For instance, in Chapter 4, this study discovered that bot-generated cascades are highly influenced by specific topics. However, the sheer volume of textual data being examined made manual coding for contextual analysis impractical. To address this challenge, this study adopted bi-term topic models[228]. These models are instrumental in analysing and autonomously categorising extensive text corpora using non-supervised machine learning methods—an approach widely embraced in existing studies of online communications[99, 229, 216].

These models are based on word co-occurrences, meaning they learn topics by modelling patterns of word-word co-occurrences[230]. Bi-term topic models have proven effective in extracting 'meaning'[99] or 'themes'[228]. They have outperformed traditional topic models like Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) in unsupervised classifications of short text, such as tweets[228, 229].

The topic modelling process consisted of three steps: (1) preprocessing the text data, (2) training and fine-tuning multiple topic model algorithms, and (3) applying the model to the corpus. During preprocessing, this study removed stopwords, usernames, emojis, and links from the tweets. Additionally, each word was lemmatised and stemmed. For stopwords, this research used an open-source list from the *nltk* package[231] designed for natural language processing, supplemented with Twitter-specific stopwords, including 'RT,' 'https,' 't,' and 'co.' Next, bi-term topic models were trained on the preprocessed data using the *bitermplus* package[232]. Consequently, two bi-term topic models were set up for all XR and BLM-related tweets, respectively. As around 70% of the tweets in the Twittersphere were retweets, only the unique tweets were included after preprocessing for topic modelling to better capture emerging trends in discourse.

These models classified XR tweets into 8 different topics and BLM tweets into 7 different topics. The determination of the number of topics was based on

both quantitative metrics and qualitative analysis. This study used the u_mass coherence score metric, whose formula is as follows[233]:

$$C_{UMAss} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}$$

The u_mass coherence score was selected due to its strong theoretical foundation in measuring the interpretability of topics based on word co-occurrence statistics[234]. While there are alternative coherence metrics such as c_v coherence[235], this study finds u_mass appropriate as it relies entirely on internal word co-occurrence statistics rather than external data sources[234], making it well-suited for unsupervised topic modelling where pre-existing knowledge or domain-specific word embeddings may not be available. Additionally, u_mass aligns well with probabilistic topic models, including bi-term topic modelling, particularly when applied to large social media datasets, where data are often noisy, unstructured, and highly dynamic[236].

It is true that other metrics such as likelihood and exclusivity could also be adopted to evaluate topic quality. Likelihood measures how well the model fits the data by evaluating the probability of the observed words given the estimated topics, making it useful for assessing overall model performance. However, high likelihood alone does not guarantee interpretable topics[237]. Exclusivity assesses how distinctly a topic' s words appear in only that topic, ensuring that topics are not overly similar to one another[238]. Although these metrics are informative, coherence was prioritised in this study due to its direct relevance to the interpretability of the topics, particularly in contexts where human analysis plays an important role in evaluating topic quality[239].

Additionally, coherence scores serve primarily as a guideline rather than an absolute determinant of topic quality. Given the social and political nature of the dataset, human interpretability played a crucial role in topic selection, which is a common practice in contexts that requires domain expertise such as social sciences[240]. Several topics with diverse keywords and content were omitted, as it was challenging to extract meaningful opinions or content from them. The

topic modelling results and analysis based on these models are further discussed in Chapters 4 and 5.

## 3.5 Approach to identifying potential causal relationships

To discern and quantify the influence of social bots on human user behaviour, this study has formulated a structured research design pipeline tailored to the research topic. Most existing studies that investigate the micro-level impact of bots on human behaviour occur within 'artificial' settings, such as computer simulations, laboratory experiments, and, at their most sophisticated, field experiments[241, 242, 243, 129]. However, this research makes use of empirical data on bot activities in real-life scenarios. This approach led to the development of a practical and replicable research design pipeline for analysing the micro-level impact of social bots. This pipeline serves as the foundation for both Chapters 4 and 5. Given that these chapters address distinct dimensions of bot impact, context-specific research designs and robustness checks have been developed and tailored to each chapter's objectives. These designs are explained in detail within the respective chapters. This research design strengthens the study's ability to assess potential bot influence and also shed light on future empirical studies of individual-level bot impact methodologically.

The research design is divided into two primary stages. In the first stage, the focus is on examining the immediate effects of social bots on human users. This is due to the bursty nature of most bot activities, which occur in waves on a large scale within a short time period, relative to the longer duration of political events. Additionally, it is common for bots to be disposable, used as one-time instances[13, 94]. Within this initial stage, this study performs the collection, pre-processing, and comprehensive analysis of the complete sample, thereby determining the chosen unit of analysis. Subsequently, suitable statistical methodologies are applied to summarise the instantaneous impacts of bots on human behaviours. In the second stage, this study begins to explore the enduring

effects of these bot activities. Throughout both stages, this study adheres to the procedural guidelines outlined by Lundberg et al.[244] to establish a theoretical estimand (the ideal measurable impact) and then link this to an empirical estimand (the feasible measurable impact) derived from the available dataset. This holistic approach allows for a comprehensive assessment of both the immediate and long-term consequences of social bot activities on human behaviour.

However, undertaking the second stage is particularly challenging due to the short lifespan of social bots and the significant inherent diversity within both bot and human groups. Moreover, data derived from real-world bot interactions, both prior to and subsequent to bot involvement, often have limitations and is susceptible to the influence of both measured and unmeasured confounding variables. In this complex scenario, the research framework follows a three-step process to investigate potential causal relationships:

a) Defining 'bot impact' and determining the sample: In accordance with the research objectives, a concise and pragmatic characterisation of 'bot impact' is established. This definition considers an evident and measurable attribute that reflects human exposure to bots on social media platforms. This defined trait acts as the criterion to filter the sample. For instance, in Chapter 5, the sample with bot impact consists of user nodes that have established edges with bot nodes within the communication network. Potential matching users include those who were active during the same period but did not meet the criteria for 'being impacted by bots.'

b) Matching users with and without bot impact: Using a matching-based design, a control group is established that closely resembles the sample with bot impact. This study then adopted common matching techniques, such as Euclidean distance between user profile metrics or Coarsened Exact Matching (CEM) designs, to minimise differences between the sample and the matching group.

However, a key limitation of this approach is that matching was based primarily on observable behavioural traits, such as post frequency and account age. While we also controlled for pre-existing political opinions towards the protest, unmeasured factors—such as cognitive traits and personality—remain potential confounders.

These unaccounted variables may introduce bias, limiting the strength of our inferences. Despite this, the study still effectively differentiates between users who interacted with bots and those who did not, demonstrating consistent behavioural differences post-interaction. While these findings are informative, future research would benefit from additional controls or experimental designs to further strengthen causal claims.

c) Further data collection and statistical inference: After creating the matching group, additional data collection is initiated as required by the theoretical estimand. With statistical inference methods such as the Difference-indifferences method in Chapter 4, this study then quantitatively assesses the impact of bots on human behaviour. These methods enable a more precise estimation of the effects associated with bot interactions.

Our key statistical inference technique is Difference-in-Differences (DiD), a method commonly used to estimate causal effects by comparing changes in outcomes over time between a treatment group and a control group[245]. In this study, the treatment is defined as interaction with social bots, and we compare behavioural changes before and after bot interaction while controlling for trends in the control group. To ensure the validity of our DiD approach, we carefully address its core assumptions, particularly the parallel trends assumption (refer to Chapter 4 Methodology). This assumption requires that, in the absence of treatment, the treated and control groups would have followed a similar trajectory. We verify this through robustness checks and statistical diagnostics. Additionally, we account for potential heterogeneous treatment effects by considering pre-existing differences in user characteristics and protest engagement.

However, this approach has its limitations. Apart from the matching design limitations previously discussed, one key issue is that we cannot be certain whether users in the control group had interacted with bots before the protest period, which may have already influenced their behaviour. This introduces potential bias, as prior bot exposure could have long-term effects that are not fully captured in the pre-treatment period. Furthermore, there is the potential impact of other human

users in the users' network: exposure to bot interactions could indirectly influence users in the control group through network effects, if they know users in the sample. These challenges highlight the need for further research using experimental or quasi-experimental designs to strengthen causal claims and better isolate the effects of bot interactions.

Despite these limitations, this three-step process still enables this study to overcome the challenges posed by the multidimensional and disruptive nature of bot impact, providing quantitative insights into the potential effects of bots on human behaviour.

## 3.6 Conclusion

In summary, this dissertation project has designed a robust, replicable, and scalable procedure tailored to the investigation of social bots and their influence on social media. The methodological pipeline encompasses all dimensions of social bot research, from initial data aggregation to sophisticated bot detection, AI explainability, natural language processing, and and approaches to identifying potential causal relationships. These aspects represent key methodological considerations in the study of social bots. While prior works have individually assessed existing methodologies and proposed their own approaches across one or more of these domains, this dissertation uniquely combines and synthesises them into a comprehensive pipeline designed for the systemic analysis of social bots.

This chapter, in particular, contributes to the existing body of research on bot detection. The identification of social bots has been a central focus since the early days of social bot studies. While tools like Botometer and various customised bot identification techniques have been developed, they have faced ongoing debate and criticism, particularly regarding issues such as false positives and overall effectiveness[36, 72, 37]. Additionally, the definition of a 'social bot' remains ambiguous within current discussions on bot detection, with only a limited number of studies addressing the conceptualisation and classification of bots beyond the simplistic human-bot binary taxonomy[246, 247].

This study contributes uniquely to the field by addressing ongoing debates and gaps in the literature through the development of a self-designed bot identification pipeline. this study conducted a thorough evaluation of existing bot detection methods and proposed a methodology that integrates various strategies to enhance the robustness of detection results. This approach addresses key concerns such as false positives and methodological reliability by incorporating rigorous robustness assessments, which strengthen the credibility of classification outcomes.

Even in the context of advanced LLM-powered bots, which have further complicated the bot detection problem, the pipeline remains effective. By refining and combining established detection techniques, this methodology not only improves classification accuracy but also contributes to a comprehensive operational definition and taxonomy of social bots. (Also refer to Chapter 6 for further discussion and qualitative investigation of social bot taxonomy). This fills a critical gap in the literature and offers a more fine-grained understanding of bot behaviour, making it a valuable tool for navigating the increasingly sophisticated world of social media interactions.

In summary, this dissertation project advances the fields of bot definition, detection, and classification by presenting a methodological approach that not only deepens our understanding but also lays the groundwork for future studies in the domain of social bots.

# 4

# The impact of social bots on human behaviour in protest communication

## Contents


**4.1 Introduction . . . . . 78**
**4.2 Results . . . . . 82**
4.2.1 Proportions of bot and human communication . . . . . 82
4.2.2 Temporal co-dependence of bot and human tweeting . . 83
4.2.3 Predicting the impact of direct bot interactions . . . . . 85
**4.3 Limitations . . . . . 90**
**4.4 Discussion . . . . . 92**
**4.5 Data and methods . . . . . 95**
4.5.1 Data . . . . . 95
4.5.2 Bot Identification . . . . . 95
4.5.3 Topic Modelling . . . . . 96
4.5.4 Testing time-series relations . . . . . 97
4.5.5 Sentiment analysis . . . . . 98
4.5.6 Matched sample . . . . . 98
4.5.7 Identifying cascades . . . . . 100
4.5.8 Support group categorisation using ChatGPT . . . . . 100
4.5.9 Difference-in-difference models . . . . . 101

# Preamble

This chapter addresses RQ1: How do social bots influence human communication behaviour in protest-related discourse? It specifically examines two impacts of bots on human behaviour: posting frequency and communication sentiment, using Extinction Rebellion as a case study. By doing so, it uncovers the micro-level effects of bots and establishes a association between bot activity and changes in human behaviour.

Structurally, this paper builds on the pipeline from Chapter 3 and addresses the research gap suggested in Chapter 2 (specifically 'protest and social media - information dissemination' and 'cases - Extinction Rebellion'). It contributes to RQ1 by demonstrating that social bots can significantly alter public discourse by changing sentiments and behaviours, and adds to our understanding of AI's societal impact and political communication. Also see appendix A for supplementary information to this chapter.

**Publication status and author contributions:** A version of this chapter is forthcoming at *Scientific Reports*. It is co-authored with Orsolya Vasarhelyi and Balazs Vedres, with the candidate, Linda Li, as the lead author responsible for the major contributions. (See Appendix D for signed author contribution declarations.)

## 4.1 Introduction

Social media has become the primary channel to engage in political discussions, and to participate in collective action over the past decade[152, 75, 5]. However, on social media there are automated agents as well, among them 'social bots,' that are increasingly active in our social and political publics[2], creating content and interacting with humans with ever increasing sophistication[248]. This hybrid ecosystem where algorithmic agents and humans co-exist can fundamentally alter the nature of democracy, political accountability, transparency, and civic participation[59]. Automated accounts can propagate a large volume of messages at minimal

expense, and can engage with users with a fast reaction speed[15]. As a consequence, the algorithmic share of social media communications is now on par with human participation: Automated users were estimated to be responsible for generating 10% to 40% of tweets in recent political events, such as the 2016 US presidential election, the Brexit referendum, the yellow vests movement, the Catalan referendum, or the 2019 United Nations Climate Change Conference[13, 14, 23].

Social bots – especially those intending to mimic human behaviour – can disrupt online political discussions[249], and significantly influence political debates and activism[249, 12, 128]. Such bots are frequently designed to pass as human accounts, and occasionally also mimic known political figures and government accounts to gain the attention and trust of human users[33]. Bots are often highly active during the flare-up of discussions around new political events[24], and disseminate targeted messages ranging from fabricated news to contentious, divisive, and negative content[18, 100, 21], blending legitimate messages and misinformation. Furthermore, bots are also often deployed in orchestrated efforts to generate the facade of a seemingly vibrant discussion conforming with hidden agendas[65, 250] by retweeting each other (known as 'astroturfing'), targeting susceptible users[13].

Although social bot presence has been studied before at the macro scale, less is known about the micro-level impact of human-bot encounters on subsequent human activism. Research on bot-human interaction found that bots can often hijack the topics and overall tone of human discussions[23, 18], and they often increase the visibility of extreme views[13, 18, 97], influence sentiment around topics[34], and intervene in human communication flows[103, 241]. Small-scale simulations and experiments indicate that bots can alter expressed human values[241] and behaviours[242], particularly driving users towards more extreme viewpoints[243] and potentially silencing them[129]. However, there is little empirical evidence regarding the capacity of bots to modify human behaviour in real-life political communication[100, 65, 103, 94].

Today, online activism constitutes an essential part of democracy. Online activism can be defined as the strategic use of digital communication technologies

by individuals or groups to engage in political and social change efforts. These efforts include social media campaigns, online petitions, digital protests, and the diffusion of messages through digital platforms to mobilise individuals or communities around specific causes, and advocate for policy changes[251]. However, most research on social bots focuses on automated interventions in institutionalised political processes such as elections[34, 252, 253, 97], while few studies examined the role of bots within activism related to online protests[23, 103, 57]. The dynamics of social media use in online activism differs markedly from communication around other political events[113, 254, 25, 26]. Since social bots are designed to be more responsive than humans[255], it is common to observe an increase in bot activity during the peak of heated online debates, followed by a decrease in their presence[13]. While the bursty nature of social media communication of protests is well known[256, 257], the impact of social bots on human activity during and after bursty periods has not been investigated.

How does interacting with social bots impact human behaviour in online activism? We address this question with data on Twitter discourse on climate-change-related social movements. We analyse the dynamics of bots and humans engaging with each other, and we also compare the difference in impact of direct communication with bots to activists who did not directly interacted with bots. We focus on events of direct communication with bots – when humans engage with bot messages in writing (replying, mentioning or commenting bot tweets) – and not merely events of seeing bot communications, as bot messaging today is highly common, and seeing one automated message should not have much impact. Our analysis focuses on protest-related discourse during a series of protest events that erupted from November to December 2019. We decided to concentrate on online activism related to climate change as our case study, as algorithmic threats to engagement in climate change activism can have profound consequences on societal agreement on the public good in a critical issue[258]. We analyse communication around the Extinction Rebellion (XR), as the highest profile activist group online.

The topic of climate change has been shown to attract highly engaged, active, and committed participants[8], while there is also substantial bot activity[42, 43]. This enables us to analyse the impact of human-bot interactions on humans during information cascades[65], and measure the effect of bot-human interactions on tweeting activity[129] and sentiment[243]. Our findings contribute to the growing body of research on machine behaviour[59], particularly in terms of understanding how rapidly developing hybrid human-machine systems could potentially modify human opinion over an extended period.

Overall, we have found that social bots are prevalent in the context of XR Twittersphere. There are significant information flow between social bots and humans. The quantity and direction of information flow between humans and bots are highly topic-dependent. However, bots were creating information cascades related to specific protest incidents online, and their activity level could be used to predict that of humans. Furthermore, compared to those who didn't interact with social bots, users who interact with bots showed a negative change in their sentiment towards the XR protest. The type of bot matters for the impact on the users' tweeting activity, and the initial level of support a user has toward the climate-change movement determines how a bot encounter impacts their sentiment on climate change.

Early work related to human-bot interactions on social media found that user sociability and network size predict who will be interacted by a bot[259, 24]. Our results show that bots have become so widespread that it is unlikely that an active user of X will *not* meet a bot online. Our research extends previous work by focusing on quantifying the impact of direct human-bot communication. These results have important policy implications for increasing platform transparency in how they handle automated profiles.

## 4.2 Results

### 4.2.1 Proportions of bot and human communication

We found 48% of all accounts to fall into the bot category within our sample from Twitter (44, 121 of the total 93, 499 users). We identified automation through a combined approach[260, 72] that integrates the results of commonly used bot identification methods, Botometer[35] and our self-trained machine learning-based algorithms. Botometer is a widely adopted open source tool to identify bots on Twitter[36]. Our self-trained models used various data sets tailored to identify users with automated behaviour that attempts to mimic humans on social networks, especially in the context of political behaviour[35]. Bots reported in the main text are the combined results of these two models, with a fixed threshold for the Botometer ($CAP >= 0.65$). Since the concept of 'bot' encompasses varying degrees of automation, using one fixed bot classification threshold is always a simplification. Therefore, we repeated all analysis reported in the main text at various bot thresholds, and we report these results in the supplementary information ('SI', appendix A).(See Data and Methods, Bot Identification and SI Bot Identification for more details. SI table A.3 shows details of the training set and data used for bot identification, while Figure A.1-A.3 shows metrics used for fine-tuning bot detection models).

Figure 4.1 shows the information flow between bots and human users. 81% of tweets were replies or retweets in our database, and 51% of these retweets originated from bots. In general, bot activity is mainly the posting of original messages or the retweeting of each other (35% of all retweets and 71% of bot retweets were retweeting other bots), and only a small portion of bot retweets were retweets originating from humans (29%). At the same time, humans spread roughly the same number of messages produced by bots (54%) and humans (46%). If the bot threshold is increased to $CAP >= 0.75$, still 48% of the retweets originate from bots, and 45% of the human retweets originate from bots (see SI Information flow and SI Tables A.6 to A.8 for information flows).

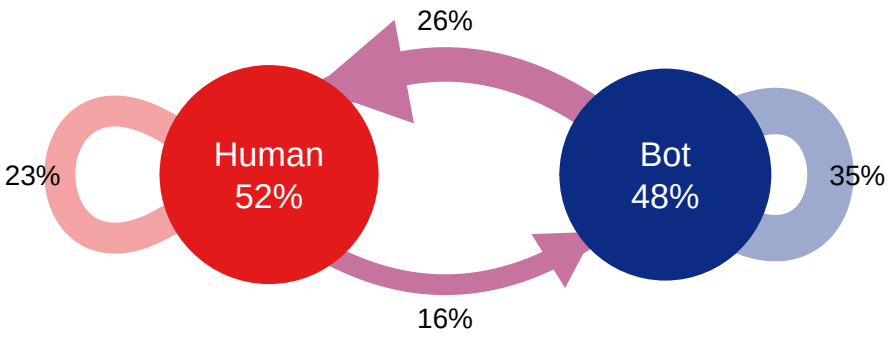



**Figure 4.1:** Information flow of humans and bots. The directions and amount of tweets retweeted between bots and humans.

It is worth mentioning that all statistical inferences, including predictive modelling (time-series based inference) of information flow (i.e., who influences whom in terms of tweet counts), were conducted using raw tweet counts rather than percentages. However, in this subsection, we primarily report percentages in the descriptive analysis to provide standardised comparisons across different categories and avoid misinterpretations arising from variations in sample sizes.

For instance, while reporting the information flow patterns between bots and humans, using percentages allows for meaningful comparisons between bot-generated and human-generated content without being skewed by differences in total tweet volumes across categories and time periods. Additionally, since the share of bot interactions in human retweets and vice versa is a proportion-based relationship, percentage values more effectively highlight how information flows between these groups. Hence, we report information flow patterns in percentages while conduct all statistical inferences in raw numbers.

### 4.2.2 Temporal co-dependence of bot and human tweeting

We found that the quantity, intensity, direction and the sentiment of the information flow between humans and bots are highly topic-dependent. We identified seven topics related to XR protests associated with news events, political campaigns, or outbursts of sentiments (for example, climate change denial) by applying bi-term topic models[228]. (See Data and Methods, Topic Modelling for more details and SI

Table A.5 for a brief summary of the topics, including their content, top hashtags, and sample tweets). Out of these seven topics, four were highly bursty: Topic-related activity increased sharply periodically and then decreased suddenly[261], resulting in information cascades. Information cascades occur when users follow the behaviour of other individuals on social networks (e.g. retweeting the same message)[262]. (See Table A.9 in SI on Burstiness Scores).

Figure 4.2 shows the temporal distribution of human and bot tweets aggregated at a 5-minute interval within an illustrative bursty period of climate change discussions on Twitter. Although human users (red line) generated a higher volume of tweets during the cascade's peak, the tweeting frequency trend is influenced by bots. Table 4.1 shows the results of Vector Error Correction Models (VECMs). VECM is a statistical model used to analyse and estimate whether one time series could be used to predict another after introducing a time lag and potential confounds [263].

We found that in 3 out of 4 identified cascades, the number of bot tweets during bursty periods could be used to predict human activity, and the sentiment of tweets by humans could predict the sentiment of bots' tweets. Bots impacted human tweeting activity significantly stronger in the case of 'Disruptive engagement' (Wald=10.06, p=0.0015) and 'Anti-XR protests' (Wald=45.36, p=0.000), but tweeting activity related to 'Politicised Activism' was mutually driven by both bots and humans. (Wald=0.76, p=0.39). Humans impacted bots' sentiment significantly more negatively in the case of 'Anti-XR protests' (Wald=3.66,p=0.048). In the case of 'Politicised Activism', sentiment was also mutually driven by both bots and humans, with no significant difference in coefficients strength (Wald=1.65,p=0.198). The effect remains consistent across bot CAP thresholds ranging from 0.50 to 0.75 for 3 of the 4 cascades identified. (See more details on Burstiness in Data and Methods, Identifying cascades, SI Cascades, SI Table A.9-A.20 for Results across different CAP thresholds, different time lags, for full time period of our analysis instead of cascades, and for Topic Burstiness Scores).

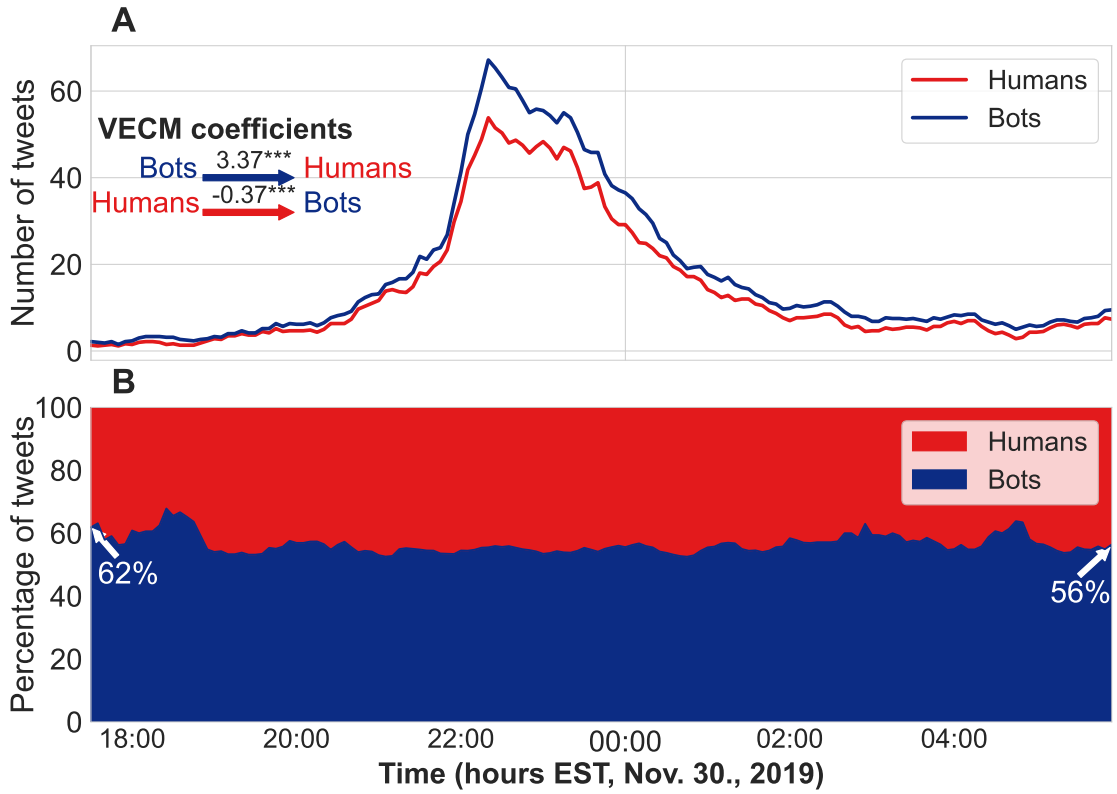



**Figure 4.2:** Panel A: Number of tweets posted by bots (blue line) and humans (red line) in a cascade mutually-driven by bots and humans within the 'Anti-XR protests' topic. Panel B: Ratio of bot and human generated tweets throughout the time period of the cascade. The number of tweets is the rolling average aggregated on 5-minute intervals. Bots are classified as users who has a overall bot probability no less than 0.65.

### 4.2.3 Predicting the impact of direct bot interactions

To test whether direct communication with bots (replying, mentioning or commenting bot tweets) has any influence on human users - beyond their above shown immediate collective effect in bursty periods - we developed two models focusing on 1) Amount and 2) Sentiment. Specifically, we analysed how human communication evolves over a span of 30 days after the first direct interaction with a bot in the discussion related to climate change on Twitter. The first model captures the inclination to 'speak out' quantified by the average number of tweets posted related to the XR protest. The second model predicts the change in sentiment about the climate change protest (measured on a scale ranging from $-1$ to 1, with $-1$ representing the most negative sentiment and 1 the most extreme sentiment).

| Topic | Amount Bots→Humans | Amount Humans→Bots | Wald test | Sentiment Bots→Humans | Sentiment Humans→Bots | Wald test |
|---|---|---|---|---|---|---|
| 'Football game protests' | -0.14 | 0.06 | 1.72 | -0.01 | 0.13 | 2.05 |
| 'Disruptive engagement' | 57.89 *** | 0.01 | 10.06 ** | -0.04 | 0.34 ** | 2.70 |
| 'Anti-XR protests' | 3.37 *** | -0.38 *** | 45.36 *** | -0.08 | -0.65 *** | 3.66 * |
| 'Politicized activism' | 0.54 ** | 0.17 *** | 0.76 | 5.09 *** | 4.12 *** | 1.65 |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table 4.1:** Vector error correction models (VECM) testing the relationship between the amount and sentiment of bot and human communication with 20 minutes time lags. VECM-s were used to determine whether bot activity predicts human activity (Bots to Humans) or human activity predicts bot's (Humans to Bots) in the identified four cascades. Wald tests were performed to test whether coefficients of Bots to Humans and Humans to Bots predictions are significantly different. Modelling were based on 30-minute time lags.

To quantify the impact of direct bot interaction on human communication change, we used a three-step process. This includes: 1) we sampled $N = 303$ users who directly interacted with bot accounts (bot-exposed) - replied or commented to a tweet/comment originated from a bot; 2) we collected a matched sample of $N = 184$ users, who were active in the XR related protest discussion on Twitter but had no direct interaction with bots. Matching users were selected on the basis of the similarity score calculated pairwise to our original sample considering publicly available metrics on human users' profiles. (See Data and Methods, Matched Sample for more details.) 3) Figure 4.3 shows two example timelines, one human user who directly interacted with a bot (the interaction is shown at time 0), and a matched user that did not encounter a bot.

3) Finally, we applied Difference-in-Difference (DiD)[245] regression models to quantify the casual effect of bot interaction on outcomes by comparing a set of humans who directly interacted with bots (human replying or commenting bot tweets) with those who did not. Our observation units are the daily activity of human users relative to the time of interaction with bots - 30 days *before* and 30 days *after*. This setting allows us to quantify the prolonged impact of bot

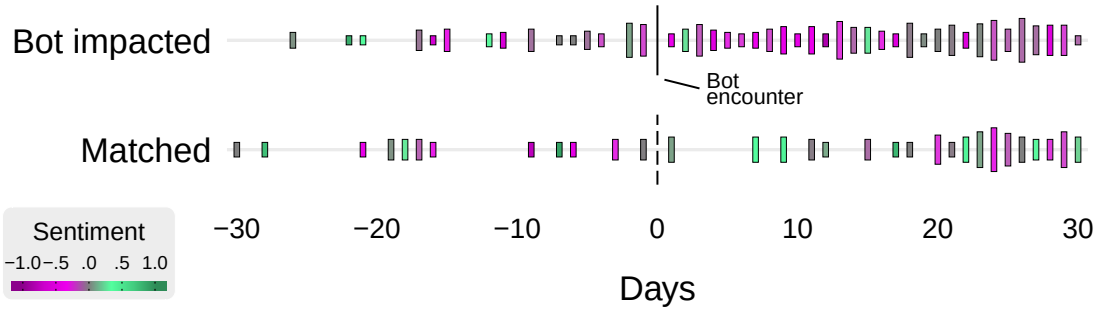



**Figure 4.3:** Two examples of Twitter timelines. Top timeline shows daily tweet counts and daily mean sentiment for a human account that directly interacted with a bot on day 0; bottom timeline shows a matched account without a direct bot interaction. Bar heights are proportional to tweet counts, color indicates sentiment from −1 to 1.

interaction on the frequency of tweeting and sentiment of tweets compared to users who did not meet a bot.

Botometer provides prediction values for belonging to seven subbot categories ranging from 0 to 5. We classify a bot into a subcategory if its bot-specific probability is greater than 2.5. Our biggest group is 'miscellaneous other bots' which are bots that are similar to various type of manually annotated bots (23%), followed by manually labelled political bots, so-called 'astroturfs' (14%),'fake followers' bots purchased to increase follower counts (7%), 'self declared' bots from botwiki.org (5%), 'spammers' (2%) and 'financial bots' (0.4%). Although 'miscellaneous other bots' are quite well represented within the XR discourse online, only 17% of human users had direct interaction with them. As implied by the topic of our analysis, the individuals in our sample primarily engaged with 'astrotufs' (38%). These automated accounts are specifically designed to participate in political discussions[75], leading us to analyse users interacting with astroturfs separately from those who interacted with non-astroturfs.

Figure 4.4. visualises the predicted daily change in the average number of daily tweet counts (Panel A), and the sentiment of the tweets (Panel B) grouped by different bot probabilities (65,70,75) and subbot category (astroturf or other type of bots). We found that interaction with astroturf bots results in an increase in the number of tweets, while communication with other kind of bots result in a

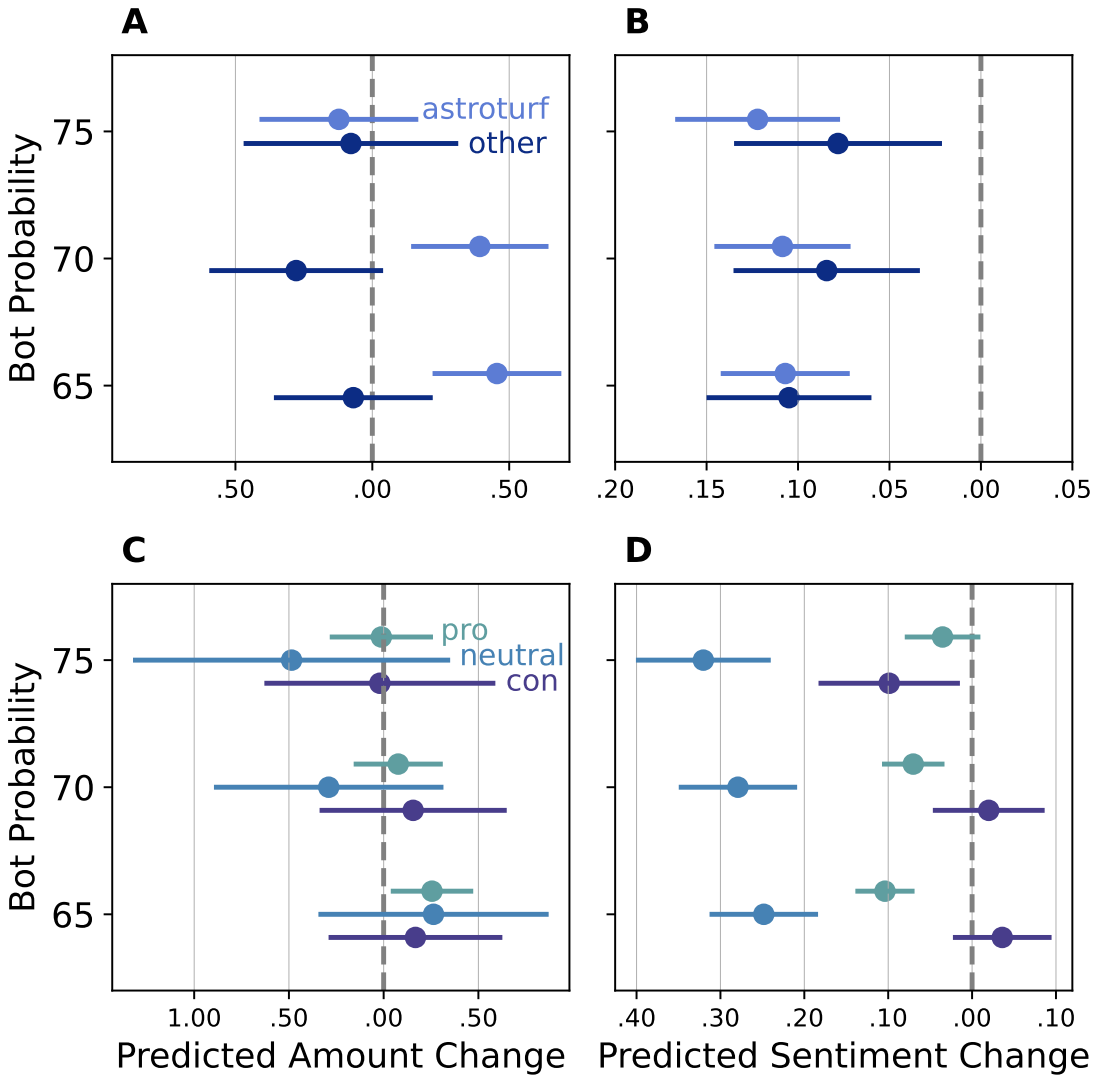



**Figure 4.4:** Predicted change resulting from a bot interaction. Panel A, C: Predicted change in Amount (number of tweets). Panel B,D: Predicted change in Sentiment. Panel A-B separate predictions (by color) are shown for astroturf bots (light blue), and other bots (dark blue), while Panel C-D separate predictions by users support level towards XR - supporters (green), neutral users (blue), anti-XR (purple)

decrease in the number of tweets. This suggests that the most politically relevant bot category–astroturf bots–drive the conversation by provoking engagement from human users, while other kinds of bots have a rather negative, silencing effect. Regardless of the bot type, direct interaction with a bot decreases the average sentiment of human users. (See SI Table A.22-A.27 for DID models)

Astroturfs tend to be mobilised in a targeted way against users with a specific opinion[24, 69]. Therefore, we classified bot-exposed and matched human users

by their support of XR Protests: Supporters (Pro - 52%), Neutral (27%), and Anti-XR (Con - 21%) using ChatGPT. (See Data and Methods on Support Group Categorisation). Figure 4.4, panel C-D indicates that human users who support or have a neutral opinion about XR are significantly affected by interacting with bots, while anti-XR users are not affected. Bot interaction has the strongest negative impact on the sentiment of bot-exposed users with neutral opinion, indicating that bots might target those users whose opinion can be changed[24, 69]. The change in the number of tweets was significantly increased by bot interaction for XR supporters within the least selective bot probability category (65), although the trends are similar in the more selective categories (.70,.75). In these models, we control for the astroturf score of the interacted bot, which has a strong positive significant relationship with the change in amount. Our results indicate that the type of bot matters more for the activity change than the original support towards protests. However, the level of sentiment change depends on the original support level of users exposed to bots. (See SI Table A.28-A.33 for DID models)

We also investigated whether bot-affected users alter their levels of support as a result of bot interaction. There were no significant differences between the distribution of bot-exposed users' support level before, and after interacting with a bot (Mann-Whitney U=45536.00, p=0.86).There was no significant difference either between users exposed to bots and those matched in terms of the change in support level.(See Model Tables A.34-A.35 in SI on Opinion Change). Out of the users who interacted with bots, only 9% experienced a shift towards neutrality, while 10% of the opinions of the users in the matched group became more neutral towards XR protests. Additionally, 7% of the users exposed to bots and 6% of the matched users changed their opinion completely, shifting from negative to positive or the other way around.

The effect of bot interaction on human behaviours remains statistically significant after controlling demographic variables (location of users), for retweets to news media reports, and with a sample matched with different matching methods (See Model Tables A.36 to A.39 in SI on other robustness checks).

## 4.3 Limitations

There are five potential limitations to our current research design. The first is the relatively short time range and sample size of our dataset on XR protests. Our dataset covers several waves of XR protests within a month, but XR is a global phenomenon that has been going on for several years. However, comparing our sample size to previous studies on online activism and political communications on Twitter[25, 113], we believe that it is a valid and representative case to illustrate the studied aspects of bot activities on online activism. The Twitter academic product track API (available at the time of data collection) provided the full archive of tweets based on specific search queries; therefore, our dataset is a comprehensive sample of the online record of bot and human activities during the protest period.

The second limitation concerns our binary bot detection method. It has three main drawbacks, each of which we address through additional efforts and measures. First, similar to other bot identification approaches[35, 215, 23, 76], and due to the nature of unsupervised learning, our method cannot be a 100% sure whether a user classified as a bot is genuinely a bot. Second, the concept of 'bot' encompasses varying degrees of automation among Twitter users, and using a fixed bot classification threshold overlooks these nuances. We are aware that being a bot, similar to the concept of gender[264], is not a binary classification problem. Many human users apply automation to increase their efficiency[246], which does not turn them into bots, but they are no longer non-automated human users. Third, a fixed threshold-based approach can still lead to false positives and false negatives, potentially undermining the validity of our claims regarding bot activities. To address those concerns, we ran our models with varied CAP scores. We also performed a series of robustness checks involving various thresholds for all of our analysis and reported them in the main text. (See Discussions, SI Information flow, Cascades, and Difference-in-Difference regression results).

The third category of limitations is due to the automatized methods used to label users' support towards XR protests and sentiment of the collected tweets. We are aware that these methods are not perfect.[265].Therefore, we validated these

results by two manual coders based on a subset of tweets and users timelines. We found that ChatGPT produced support level values correlated highly with human coders ($corr = 0.88$). Based on assessing the accuracy of the VADER algorithm used for sentiment analysis, we found that it categorises slightly more tweets as neutrals than human coders did. However, we found that the results are sufficiently accurate in all sentiment categories. (See Precision, Recall and F-score by sentiment category in SI, Table A.21)

The fourth limitation concerns our Difference-in-Difference sampling design. We collected a comprehensive archive of tweets during the sampling period to ensure that our control (matched) group did not interact with social bots during that time. However, we cannot definitively confirm whether matched users had interacted with bots prior to the sampling period. Such limitation introduces a potential issue for our claims of bot impact, and necessitate careful interpretation of the relationships.

However, our findings show that orchestrated bot activity was highly concentrated and bursty around key protest-related events, suggesting that bot influence was reactive to these news peaks. It is unlikely there was much sustained bot activity before the outburst of protest events, which implies that any potential effects from earlier interactions would be minimal. Furthermore, based on our observation of changes in sentiment and retweeting behaviour, the effects of bot interaction appear to be short-lived. As a result, any bot interaction that occurred prior to our sampling period would likely have diminished by the time our analysis begins. Therefore, it is unlikely that pre-sampling bot interactions would significantly impact our estimates.

Furthermore, our analysis already reveals significant differences in behaviour between users who interacted with bots and those who did not during the sampling period. If some control group users had interacted with bots before the sampling period,we would expect them to display similar behavioural changes, like shifts in sentiment. In this case, both groups would have users who have experienced bot-induced behaviour changes. As a result, the observed differences between the

two groups would become smaller than the actual effect, making our estimates of the treatment effect more conservative.

The final limitation concerns the potential influence of unmeasured confounders in our Difference-in-Differences (DiD) design. While we have accounted for several confounding factors, there may still be unobserved variables that affect the differences observed between the treatment and control groups. For instance, individuals in our sample may have been more likely to interact with bots due to pre-existing personality traits, prior attitudes toward XR or environmental activism, or varying levels of awareness regarding the prevalence of social bots. These factors could introduce self-selection bias, which may impact the interpretation of our findings.

However, the primary goal of this study is not to establish a definitive causal relationship but to demonstrate that individuals who interacted with social bots exhibited behavioural changes, such as shifts in sentiment and posting frequency, compared to those who did not interact with them. This pattern holds true regardless of whether individuals actively chose to engage with bots or not. The fact that bot exposure is associated with measurable behavioural differences already carries significant implications for understanding the role of social bots in shaping online discourse.

Nonetheless, future research is needed to strengthen suggested causal claims by addressing potential unmeasured confounders. Experimental approaches would be particularly valuable in isolating the precise mechanisms through which bots influence behaviour. Such methods could help establish more robust evidence on how bots affect users, ultimately informing more effective strategies for mitigating their impact.

## 4.4 Discussion

Bot presence is considerable and possibly increasing in the public sphere. Even with stricter thresholds, bot activity is higher in our sample than in previous related work published on the 2017 Catalan referendum by Stella et al.[18], where only 19% of all interactions were from bots to humans. It was shown that tweets from conservative

bots are more retweeted by humans, which can indicate that this difference is not only due to the continuously increasing presence of bots[71], but could be due to the highly politicised and international nature of our context.

Even though we considered only one public in detail, we replicated our descriptive analysis of information flows on similar data from Twitter about #BlackLivesMatter movement. At the active phase of exchanges from the time of the first protest followed George Floyd death, we found that the patterns of how bots impact human communication are are not significantly different in the two context (See SI Table A.1 for more details). Although XR is one of the most prominent fields of protests online, with highly committed activists which can lead to increased bot presence and more targeted actions against them, the similarity between the two cases indicate that increased bot presence is a general threat that activists face in social media.

We have adopted a dynamic approach to political communication and have found that bots should not be thought of a constant presence, but rather bots are driving the amount of heated and bursty discussions, and react in a dynamic fashion to human sentiment. Our temporal analysis of human and bot activity showed that in 2 out of 4 identified cascades bots impacted human tweeting activity significantly more, however humans drove the sentiment of the communication. We also found that bots and humans can mutually drive cascades, depending on the topics and the intensity of the debate.

To quantify the potential impact of human-bot communication, we have compared the communication histories of human users who have directly communicated with a bot with those matched human users who have not. This allowed us to see a consistently negative impact of any kind of bot interaction on the sentiment of subsequent human communications: Humans who have interacted with a bot displayed considerably more negative sentiment than matched users.

The change in tweeting activity after a bot interaction depends on the nature of the bot: On one hand, decidedly political astroturfing bots (aiming to influence public opinion behind an impression of grassroots opinion) result in an increased

activity. On the other hand, interacting with other kinds of bots (spammers, fake followers) results in a decrease in activity. However, change in sentiment towards the protests depends on the original support level of the user, supporting, neutral, or against XR. Our results indicate that bots might target users whose opinions are easier to change, since the sentiment of bot-exposed users with neutral opinion decreased the most. However, bots do not make human users switch their support level. In sum, bot interaction is not without impact, even if one encounter itself has only a small effect (it takes about two bot encounters to induce one additional tweet). Nevertheless, since there is an exceeding amount of bot communication, these small impacts add up to influence the public sphere in a profound way.

Although our analysis covers a period of time prior to the launch of ChatGPT (30.11.2022), it is becoming increasingly difficult to identify bots due to the rapid advancement mimicking human behaviour[266]. As large language models are becoming widespread and easily accessible through APIs, new social bots can act extremely human, making it currently almost impossible to distinguish between bots and humans, even for experts[92, 267].

Therefore, it is important to have unrestricted access to social media data to assess the influence and prevalence of these new types of social bots, although recent trends show that social media platforms are less willing to share free data for research purposes[214, 268]. Since financial evaluations are highly correlated with the size of the (human) user base[269], platforms have no interest in quantifying the ratio of non-human accounts and their impact on misinformation[24]. Most users still underestimate the effect of bots on themselves, but as they are exposed to increased bot presence, they tend to prefer stricter bot-regulation policies [270]. That is why we welcome the news that the European Union requires larger platforms to provide researchers with access to data to study systemic risks arising from the use of their services, such as disinformation[271]. Such legislative actions can help the scientific community continue its work to understand the consequences of this abrupt change in technology that will alter the nature of human-bot interactions[92, 272, 273, 274].

## 4.5 Data and methods

### 4.5.1 Data

Our data is made up of Twitter activity around several waves of Extinction Rebellion climate change protests from 18 November 2019 to 10 December 2019. The dataset was collected from November to December 2022 via the Academic Research product track API provided by Twitter[213], which enabled users to collect a full archive sample of historical tweets filtered based on keywords and conditions. We collected all tweets posted during this period of time that contained the keyword 'Extinction Rebellion', 'climate change protest', 'XRebellion', 'XR' and multiple variants of keywords with slightly different spelling. (The complete list of keywords used can be found in SI Table A.5) In total, the final data set contained 201,010 tweets and 122,130 users.

### 4.5.2 Bot Identification

To identify social bots on Twitter, we used a combination of two sets of bot identification methods. The first is a popular Twitter bot identification tool known as the 'botometer' (formerly BotorNot), which was primarily used as a benchmark to compare other methods. The second is a set of our self-trained bot identification model trained with open source data of bots and humans to train supervised machine learning models.

Botometer is a publicly available tool that relies on machine learning. It is designed to calculate a score where low scores indicate likely human accounts, while high scores suggest likely bot accounts[275]. The algorithm considers more than 1,000 features related to user profiles, friends, network structure, and activity patterns, among others. Another part of our bot identification pipeline comes from self-trained models. Training sets were derived from existing open-source data from Twitter accounts identified as 'bots' and 'humans'.

We trained bot identification models with 70% training and 30% of testing set with five types of algorithms: random forest(RF), support vector machine(SVM),

logistic regression(LOG), XGboost classification (XGB) and deep learning (DL). We developed two versions of our models with ten and twenty features that were proved to be most effective for bot identification by previous studies[276, 23]. The evaluation of the models demonstrated that the RF, DL, and XGB models with 20 traits surpassed other models in terms of sensitivity (true positive rate), balanced precision, precision, and F1 score. Additionally, these models exhibited strong performance in an independent test carried out on a dataset consisting of influential bots and human mimics that were active during the 2018 midterm election of the United States.

In our final bot identification approach, we combined the results derived from both sets of bot identification methods. Due to the potential false positive issues inherent in both methods, they yielded somewhat divergent results[36]. To reconcile this disparity, we classified users based on the overlapping results of botometer *and* our proprietary algorithms (DL, RF, or XGB). If both the botometer and at least one of our algorithms identified a user as a bot, the user was classified as such; conversely, the same principle was applied to the classification of humans. To account for the potential error in bot identification, we performed all our analyses with a varying baseline CAP score ranging between $0.60 - 0.75$ (See more details in Supplementary Information, Bot Identification).

### 4.5.3 Topic Modelling

We first identified the themes emerging in the protest-related discourse with the increasingly popular bi-term topic models[228, 229] that learn topics by modelling word-word co-occurrence patterns[230]. After removing all retweets, to preprocess the data we used the nltk python package[231] to remove stopwords, usernames, emojis, and links from the tweets, and lemmatise and stem every word.

We then trained bi-term topic models on our preprocessed data with the bitermplus package[232]. We set up a biterm topic model of all the tweets related to XR, which classified all tweets into 8 different topics. Based on the u_mass coherence scores[233] (See Figure A.5 in SI), we determined that 8 topics fits our dataset the

best. After evaluating the meaning of topics we dropped one topic whose keywords and content were too diverse to extract a meaningful media agenda from it. (See more details in SI, Topic Modelling).

### 4.5.4 Testing time-series relations

To test whether bots influence human activity or vice versa, we employed Vector Error Correction Models (VECM). VECM can handle multivariate time series data that are likely co-integrated, which means that they share a stable long-term relationship despite short-term fluctuations[263]. Therefore it is more suitable compared to other methods, such as vector auto-regressive (VAR) models, or granger causality tests it does not require stationary time data.

We created aggregated time series of the number of posts by bots and humans, then introduced a 30-minute time-lag between the number of tweets and the average sentiment by bots and by humans. The augmented Dickey-Fuller (ADF) test, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests and the Johansen test all suggested that all time series we tested were non-stationary(e.g., the mean or variant was not constant or there were seasonal fluctuations in the time series trend), but stationary for their first order difference. However, first-order differences of the time series are stationary, suggesting that they satisfy the prerequisite of VECM. The formula of VECM models is as follows:

$$\Delta Y_t = \alpha + \sum_{i=1}^{p-1} \Phi_i \Delta Y_{t-i} + \sum_{i=0}^{q} \Gamma_i \Delta Z_{t-i} + \beta_1 (Y_{t-1} - \beta X_{t-1}) + \epsilon_t$$

where $\Delta Y_t$ represents the vector of first differences of $m$ variables $Y_1, Y_2, \ldots, Y_m$ at time $t$. $\alpha$ stands for the constant term, $\Phi_i$ are coefficient matrices for the lagged first differences of $Y$ variables up to $p-1$ lags, $\Gamma_i$ are coefficient matrices for the lagged first differences of $Z$ control variables up to $q$ lags, $\beta_i$ is a coefficient vector capturing the speed of adjustment (error correction term), $Y_{t-1}$ denotes the lagged level of the $Y$ variables, $X_{t-1}$ denotes the lagged level of the $X$ control variables, and $\epsilon_t$ represents the error term or residual at time $t$.

Apart from our main variables of interest (bot posts and human posts), we also included control variables in our VECM models. We control for the number of retweets of major news media communication relevant for our geographical coverage (US, UK, Canada): BBC News, The Guardian, The Telegraph, The Independent, Sky News, Channel 4 News, ITV News, Daily Mail, Financial Times, Metro Uk and CNN. It is primarily news media reports that prompt activists and social bots to react and engage, thus events of news media communications can be the third variable that influences the timing of both human and bot engagement.

If the number of posts by bots could predict (statistically significant coefficients in VECM models) that of humans, then we would infer that bots were directing humans' attention, and vice versa. If no statistical significant was observed on both sides, or if the effect was mutual, following existing research practice, we conclude that bots and humans were driving the cascade together. [277, 115, 27] (See SI Cascades and Table 4.1 for detailed information of the VECM test results on all identified cascades, and further statistical robustness tests.).

### 4.5.5 Sentiment analysis

For sentiment analysis, the VADER package was used, which is an open source rule-based model and has been proven particularly effective for the classification of feelings in text from social media[278]. We used the raw text of tweets for sentiment analysis, as suggested by the package documentation. For each tweet, the algorithm assigns a sentiment score from $-1$ to $+1$, with $-1$ being the most negative, $+1$ the most positive, and 0 neutral. (See SI Figure A.6 and A.7 for the sentiment distribution of all tweets by exposed and matched users, and Section 'Validating Sentiment Analysis', SI table A.21 for details on the accuracy of the VADER python package.)

### 4.5.6 Matched sample

In order to understand how bots shape in the longer term (30 days after bot interaction) human tweeting activity and tweet sentiment related to XR, we created

a sample of matched human users who did not interact with bots in our dataset to compare with 'bot-exposed' human users. We define bot interaction as very direct and active engagement of retweeting a bot (such as replying to a bot's post), not merely seeing a post from a bot. We first identified a group of 506 users in our dataset as our exposed sample, the ones who directly interacted with bots by quoting or replying to bots. Then we calculated a similarity metric between all the 'non-exposed' human users and exposed users. Specifically, we calculated Euclidean distance based on the following metrics: statuses count, followers count, friends count, favourites count, listed count, followers growth (average number of followers increased on a daily basis), friends growth, favourites growth, listed growth, follower friend ratio. These traits were selected and/or calculated based on the user profile collected via Twitter API V1.2. The formula for the Euclidean distance is as follows, in which $p_n$ and $q_n$ means the $N_th$ trait for the sample and the matching:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

It is worth mentioning that not all exposed users have matched users, and not all matched users were active during the time window in discussions related to XR. This is because the distribution of activity levels in online political communication is right skewed[113]; half of the users posted less than five times about XR in our data set. After dropping non-active potentially matched users, in total we had 184 matched users for 303 exposed users. If a user has more than one matched user, we include up to the top five matched users in our dataset.

Our grouping passed the parallel trend assumption test (Chi-squared, p=0.076 for user sentiment, and p=0.967 for amount of posts), indicating no significant difference in the slopes of the trends for the treated and control groups before the treatment. We also applied Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM) which yield similar categories. (See SI section 'Matching Design' for details on matching-related robustness checks.)

### 4.5.7 Identifying cascades

Cascades were identified by calculating the temporal density (the percentage of tweets that belong to a given topic in a given time period) of each topic. Based on previous studies on the life cycle of information cascades online[279, 280], the time window was specified as 1 hour. To identify cascades, we identified bursty periods by calculating the Z scores of the average topic density per hour for each topic. The Z-score is a standardised measure that indicates how many standard deviations a data point is from the mean[281]. It is calculated as:

$$Z = \frac{X - \mu}{\sigma}$$

where $X$ is the observed value, $\mu$ is the mean, and $\sigma$ is the standard deviation. A high Z-score suggests a significant deviation from the average topic density, indicating a bursty period of heightened activity. This method allows for the detection of surges in topic discussion while accounting for natural fluctuations in volume. This measure was adopted by similar studies in social media [279, 280, 281].

We then filtered out all time units that had a Z score larger than 2 ($>$ 95% percentile) in any two or more consecutively two one hour time windows. Because the z score and the topic density could be high in time windows with only a few tweets, we also dropped those topics with no more than 50 tweets in at least one hourly time window. (See SI cascades for detailed information on all identified cascades.)

### 4.5.8 Support group categorisation using ChatGPT

For our research objectives, we also categorised users' opinions regarding the protests they engage in discussions about. To achieve this, we seeked to determine whether users are in favour of or against the protests they discuss. This is measured using a scoring system ranging from -1 to 1, where -1 indicates complete opposition to the protest, 1 denotes strong support for the protest, and 0 signifies a neutral stance or unrelated discussion in their tweets. Subsequently, we employ a tri-category classification scheme for further analysis: scores between -1 and -0.1 are classified

as 'Anti-XR (Con protest)', those from 0.1 to 1 as 'Supporters (Pro protest),' and scores between -0.1 and 0.1 as 'Neutral.'

The data used for this classification consists of users' tweets from our dataset. We evaluate the opinion expressed in all interactions (human replies to bots) between our sample users and bots. Additionally, for each bot-exposed user and their matched counterparts, we classify their opinions based on all tweets from their timeline before bot interaction.

We employed OpenAI' s large language model (LLM), ChatGPT 3.5[218], to classify users' standpoints. This method has been raised and adopted in various studies[219, 220]. For each user, we provide the model with an instruction prompt on how to classify their opinion toward climate change protests in general, along with the text to classify (users' tweets), and the model outputs the aforementioned score. To generate the three opinion scores mentioned above, we used the full timeline for each user before bot interaction (for opinion before), the full timeline after bot interaction (for opinion after), and human replies to bots' tweets during bot interaction (for opinion during interaction). (See SI Support group categorisation, Table A.4 for detailed information on prompt engineering and verification.)

### 4.5.9 Difference-in-difference models

We apply difference-in-differences (DiD) analysis to assess the effects of the two-way treatment on human users who directly engaged with bots. DiD is a statistical model design that incorporates both a treatment and a control group. In this approach, we estimate the effect of treatment by analysing time series data from both treatment and control groups. We compare the treatment effect of users who had direct interaction with bots and those who did not, 30 days after direct interaction.

The estimator and formula of a DiD model is as followed[245]:

$$Y_{it} = \alpha + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + \epsilon_{it}$$

In the estimator, $\text{Treat}_i$ is the key explanatory variable of differences in the treatment state, and $\text{Post}_t$ is the dummy temporal variable that says if it is before or after treatment.

For models estimating impacts on tweeting 'amount', we included days without any records of tweets (zero-tweet days) into our dataset for estimation. Because of the excessive number of zeros in the dependent variable in this case, we used zero-inflated negative binomial models to calculate the effect of bot interaction on the number of tweets. Since average daily sentiment is normally distributed, we used Linear Models to estimate bot impact.

We also control for variables that can provide alternative explanations for our findings. Our topic analysis revealed that bot activity is topic dependent and bots generate cascades, consequently influencing human communication. Throughout this process, bots might interact with humans multiple times. Therefore, we control for burstiness, the topic of the interaction, and the total number of bot interactions for each user. The sentiment of the interaction and the popularity of the original tweets could impact the level of activity in a thread. Longer threads may attract more bots, and we take these factors into account as well. An analysis of the demographic features (See SI Matching Design) of the sample and the matched groups indicated no statistically significant difference in gender. However, there was a significant difference in geographical locations. Consequently, we included location categories (UK and Ireland, Europe, USA, Australia and New Zealand, other locations) as additional control variables.

In our models that compare the impact of bot interaction by support categories, we also control the support level of the interacted bot and their astroturf score. (See SI Table A.22-A.39 for full model tables).

# 5

# The impact of social bots on online protest networks

## Contents


**5.1 Introduction . . . . . 104**
**5.2 Findings . . . . . 111**
5.2.1 Hybrid communication network during protest peak . . 111
5.2.2 Bot impact on human connection . . . . . 114
5.2.3 Moderation effects of pre-existing opinion . . . . . 119
**5.3 Discussions . . . . . 120**
**5.4 Data and methods . . . . . 124**
5.4.1 Data . . . . . 124
5.4.2 Bot detection . . . . . 125
5.4.3 Network construction and filtering . . . . . 126
5.4.4 Opinion categorisation using ChatGPT . . . . . 127
5.4.5 Topic modelling . . . . . 127
5.4.6 Community detection . . . . . 128
5.4.7 Statistical analysis design . . . . . 129
5.4.8 Individual-level analysis . . . . . 130
5.4.9 Community-level analysis . . . . . 132

# Preamble

This chapter addresses RQ2: How do social bots impact the network structure of online social networks during protests? Using a sample of social media users from the peak of the Black Lives Matter protests, it examines whether bot interactions during the protest peak have a lasting effect on human-to-human connections (network cohesion) within the online social network.

Structurally, this paper builds on the methodology outlined in Chapter 3 and addresses the research gap identified in Chapter 2, particularly in the areas of 'protest and social media–network and organisation' and 'cases–Black Lives Matter.' It contributes to RQ2 by showing that social bots can weaken human connections at both individual and community levels. This chapter enhances our empirical and theoretical understanding of AI's societal impact on online social networks. See Appendix B for supplementary information.

**Publication status:** This chapter is being prepared for submission to *Nature Communications*.

## 5.1 Introduction

In the 21st century, large-scale protests increasingly rely on social media as a central platform for organising collective actions[5, 28]. From the Occupy Wall Street movement and the Arab Spring[8] to more recent demonstrations like Black Lives Matter (BLM) and climate change protests[25, 282], social media serves not only as a means to disseminate information but also as a vital arena for political engagement. This activism predominantly occurs through the formation of online social networks (OSNs) on social media platforms, which facilitate communication, coordination, and mobilisation[113, 283] and amplify the visibility of the protests[122]. Users on social media engage in interpersonal exchanges and form social ties during protest communication[139], and therefore form distinct OSN structures online[122]. Such

connection and network structure impact information diffusion[284] and recruitment efforts[285] differently throughout the protests. Moreover, these networks can be used for purposes beyond mobilisation and awareness-raising, serving as tools for misinformation dissemination and counter-protests[286, 287, 288]. Consequently, these ties, structures and networks formed during protests hold profound implications for digital civic participation.

At the same time, automated agents, especially the so-called 'social bots', has also been attracting scholarly attention in the last decade due to their prevalence, activity, and potential impact on human networks. Empirical evidence has shown that social bots have increasingly become significant instruments of influence in previously human-dominated networks[12], and a great number of users have been exposed to content produced by bots[270, 18]. Additionally, bots attempt to establish connections with users[96] and form distinct social network structures[80] that may impact the entire human online social network (OSN). Such behaviour is particularly significant in political events over the last decade, such as the Brexit referendum[13], climate change protests[42], and Black Live Matter (BLM) protests[211]. Social bots have profound potential to impact all aspects associated with OSNs in protests. Therefore, there is a critical need to investigate the activity and potential impact of social bots within online protest networks.

There are only a few existing studies[23, 289] on the literature of bots and protest OSNs, and there are several gaps that need to be filled in the literature. The first gap is that we lack work that investigate the impact of social bots in the the context of civil engagement, especially protests. Civil activism is an important component of modern democracy[290], but existing literature on bots and social networks mostly focused on their impact on formal political processes such as elections and referendum[14, 100, 13]. Only a limited number of past work discussed bot impact on protests. However, due to civic political participation's importance to democracy, it is important to study the influences of social bots upon and within civic political participation. It is true that there are a few pieces of work that discussed bot activity impact on protests, such as the Yellow Vest

movement[23], the protest towards corruption[57] and several other protest cases in Central and South America[291, 103]. However, despite their importance, research on bots' impact on human users in the context of protests is still limited. In these studies, we first need descriptive statistics to determine whether these impacts remain consistent with previous cases or if new patterns have emerged. Therefore, we raise our first research question (RQ5.1):

> RQ5.1: What are the structure and context of hybrid human-bot online networks that emerged within discourse surrounding the Black Lives Matter protests?

The second gap is the lack of research exploring whether social bots within protest networks affect human-human ties and consequentially network structure within social network. Previous studies examining the social network structure of social bots primarily focused on identifying botnets online[76, 292] and providing descriptive accounts of their structure and metrics[80, 13]. Only a limited number of research have focused on investigating whether and how these bots influence human activities within online social networks (OSNs). These studies typically describe the structural characteristics of bots within (protest) networks, such as node centrality[23, 247]. Drawing from these structural network metrics, prior research then attempted to estimate the impact of social bots on various processes essential to protests, such as information diffusion[34, 143]. However, we still lack research that explores whether social bots within protest networks affect the maintaining of edges, as well as structural metrics of human-human ties in networks. Such metrics and movement patterns of OSNs hold significant implications for processes like protest recruitment, mobilisation, and the sustenance of civil activism communities online[28]. Filling this research gap is essential for comprehensively understanding the dynamics of digital activism within online protest networks.

The third research gap is that more studies are needed to investigate the long term effects of bot activities in network. Past work has shown that social bots are active in the peak of protests[293, 23]. Furthermore, they may remain in online social networks after the peak of protest events, with high centrality rankings,

such as the case of U.S. presidential election and Ukraine conflicts[94]. However, few studies investigated the long-term effects of bots towards human networks, and therefore we need to fill this gap. Based on the second and third research gaps, we raise RQ5.2:

> How does interaction with social bots impact the network structure of humans in the online activism context after the peak of protest events?

The fourth research gap concerns that we need a more fine-grained concept of both social bots when evaluating their impact to human users. both past work[247, 75, 35] and the discussion in our previous work[293] have suggested that the social bot is a multidimensional concept. Previous research has identified various types of social bots, including astroturfers (those orchestrating fake lively discussions), spammers, fake followers, and financial bots[75]. Furthermore, as different kind of bots have different features and strategies, they may have different impact towards human users, and it is essential to differentiate those impact. However, most past work adopts a binary human-bot classification[13, 18] that is insufficient to capture the multi-dimensional nature of different types of bots and their impact on human connections. Therefore, it is essential to identify and understand different kinds of bots and their impact to human behaviour.

> RQ5.3: How do different kinds of social bots impact human network structure differently?

The fifth gap is the heterogeneous nature of human features that was not taken into consideration by most past research of how bots impact humans' OSN structure. A few pieces of evidence from lab experiment empirical studies first find that different human users have different susceptibility to bots[294, 295]. One of the most prominent factors is the pre-existing opinion of users. users who shows opinion congruence with social bots are more likely to accept a social bots' attitude[58]. Furthermore, past research found that sentiment changes towards climate change protests depend on the user's original attitude towards the protest itself[293]. However, no existing research investigated whether such difference in

pre-existing opinion moderated bots' effect to human network structure online. Therefore, we propose RQ5.4:

> RQ5.4: Are human users with different opinion towards protest impacted by social bots differently?

To explore our overarching research topic-social bots' impact on human networks-we use empirical data from the Black Lives Matter (BLM) movement, specifically focusing on the outburst of social unrest following George Floyd' s death on 25 May, 2020. BLM is an ongoing social movement protesting against government violence, particularly targeting African American citizens. The movement gained publicity in July 2013 after the acquittal of the white police officer George Zimmerman in the death of Trayvon Martin. Our research focuses on the wave of protest events that erupted in early June 2020 after George Floyd's death while in police custody in Minneapolis, Minnesota. The incident was captured on video and widely circulated on social media. It ignited widespread outrage, demands for justice, and prompted extensive protests across the United States[46].

This case is especially relevant to our research question and effectively helps us address the gaps in past research for two main reasons. First, BLM activists extensively used social media platforms, particularly X (formerly Twitter), to disseminate information and mobilise participants for protests[25]. Given the movement's significant scale and reliance on social media, BLM serves as an ideal case study to observe and analyse the cohesion among protest participants. Additionally, past studies have indicated a substantial presence of bots within the BLM communication landscape since 2013[296, 211]. Consequently, BLM offers a valuable context to investigate bot behaviour and their impact on human connection in the digital field.

Furthermore, the BLM case is highly generalisable to the two broader contexts of this study: online social media activism and bot activity. In terms of online social media protests, the movement relies heavily on social media, particularly hashtag activism, for online mobilisation, engaging a vast and diverse range of online communities. Moreover, it is a recent, representable large-scale hashtag activism case that have gained international attention. Studying communities in

BLM provides valuable insights into how other protests make use of online social networks for organisation and discourse.

In terms of bot activity, previous research [293] has shown that BLM exhibits similar characteristics to other protests, such as XR, particularly in the way bots contribute to information flow on social media. This similarity suggests that findings from the BLM case can be extended to other protest movements where bots play a significant role in shaping online communication. Since our study focuses on the structure and dynamics of online communication networks, BLM provides an ideal case for examining how automated agents influence information diffusion within protest movements. Therefore, BLM is generalisable both in the broader context of online protests and in the specific context of bot-driven information dynamics, making it a strong and representative case for research in this context.

To address our research questions, we first construct an online community network formed by bots and humans in the case of BLM. We then identify hybrid online sub-communities with methods like community detection, which is a data-driven approach to identify nodes inside a community that are more densely connected together, to partition the network graph[297]. Doing so allows us to capture subgroups, understand network structure, and provide context to the statistical inference of bot impact on humans[113]. Furthermore, we also use methods like topic modelling to understand the main themes of discussions emerging in the hybrid OSNs, and examine how those topics are distributed across communities, as suggested by past research[298]. With a combination of topic modelling and community detection, we can get an idea of the different actors, interest groups or communities emerging in networked connective action[299]. We investigate how bot activities at the protest peak are distributed across those different online communities, as shown by features of their network and community positions and structures. By doing so, we would effectively address RQ5.1.

Our next research focus lies in understanding how interactions with social bots influence connection among human participants in protests and network structure in the long term (RQ5.2). To address RQ5.2, we examine both the immediate and

long-term effects of human-bot interactions (up to two months after the peak of protests) with our unique dataset of human interaction and connections.

To comprehensively understand the influence of social bots on human-human ties and network structure in protest online social networks (OSNs), it is essential to measure changes in human networks at both individual and community levels. This dual-level approach is necessary for both theoretical and methodological reasons.

Theoretically, ties in social network reflect social cohesion, which has varied definitions at both the individual and community levels. Social cohesion at the individual level emphasizes the strength and quality of micro-level personal connections[300, 134], while at the community level, it encompasses the overall integration and solidarity within the group[301, 302]. By measuring cohesion at both levels, we can better capture the complexity of how social bots impact both personal relationships and broader network structures.

Methodologically, past research often measures cohesion using network metrics[303, 122], and differentiates between individual and community levels to address these dual definitions of cohesion[304]. This approach provides a more comprehensive understanding of how disruptions in individual ties might affect or reflect changes in community cohesion. Consistent findings across both levels offer stronger empirical support for theories that address the diverse dimensions of social cohesion in network studies and online activism.

Our individual-level analysis focuses on the ego network. Ego network-level analysis is essential because analysis at the individual level captures the micro-level interactions between humans and bots, which affect individual perceptions of online connections[134], protest participation[305], and information diffusion[135]. These micro-level influences can then indirectly influence broader protest dynamics. In contrast, community-level analysis focuses on community identified by community detection methods. It provides a macroscopic view of how sub-groups within the larger network are organised and connected, revealing how bots impact the grouping patterns and overall network structure[122].

We aim to further investigate the impact of different types of social bots on human network structures (RQ5.3) and how they moderate the effects of bot-human interactions. In our study, we will explore various social bot classifications based on established detection algorithms, focusing particularly on astroturfers. Prior research[75] highlights the significant role of astroturfers in shaping online political discourse and their persistent influence on human behaviour within these communication settings. We will analyse whether astroturfers and other types of social bots exert differing effects on human network connections.

Additionally, we will assess how the pre-existing opinions of human users (RQ5.4) —particularly their support for movements like Black Lives Matter (BLM)—moderate the impact of bots on their network connections. By controlling for this variable, we can determine if users' pre-existing beliefs influence how they engage with and are affected by social bots. This analysis will help us understand how ideological alignment affects the interaction dynamics between users and bots, providing deeper insights into the role of bots in shaping online activism.

In addressing all those research questions, we found that bots and humans form several distinct, polarised hybrid online communities that have different view towards the BLM protest. Furthermore, interactions with social bots are associated with a decline on social cohesion, both at individual level (ego-network) and community level (sub-communities). We have also found that different bots (astroturfing and other bots) contribute to this effect differently, and pre-existing opinion of human users moderate those effects. Overall, those findings shows that bots may have a impact to social cohesion in online activism, which has important implications to studies of both social bots and online activism.

## 5.2 Findings

### 5.2.1 Hybrid communication network during protest peak

Overall, social bots and human users at the core of the George Floyd protest form a fragmented but persistent network. To construct this network, we filtered out the k-core 2 networks during the protest peak and built their communication networks

before and after the peak. The network statistics for these periods are summarised in Table 5.1. Density, clustering coefficients and reciprocity are measured on the sub-graph with human nodes only in the cluster.

At the protest peak, users formed a network with 103,050 nodes, with 55.3% (50,982 users) displaying automated-like behaviour. The number is slightly higher but still similar to estimation of bot percentages (around 40%) in past research of other protest events[23], such as the Indigenous Rights Movement. Longitudinal snapshots taken before, during, and after the protest reveal that approximately 64.4% of human users remained active within the same network, while the figure for users suspected to be automated is 78.2%. This suggests that bots are not only instrumental during the peak of the protest but also maintain a persistent presence in the network afterwards.

Even in the k-core 2 network, network snapshots are fragmented, characterised by a sparse structure with numerous isolated clusters, with low clustering coefficients for the protest periods before, peak and after. Despite this fragmentation, the protest network at its peak is dominated by a few highly influential nodes (central actors) that received a disproportionate number of reposts from different users (medium unweighted in-degree 2.0, average 6.37, maximum 12,605). This is consistent with descriptions of the core-periphery structure of online protest and political communication networks[140, 206] (see Data and Methods, network construction. Also see Table B.3 for descriptive statistics for the network snapshots).

To cluster individuals according to their communication patterns, we identified four distinct large community clusters at the protest peak using label propagation community detection methods. To further understand the topics of discussion among these groups, we conducted topic modelling on the tweets forming the edges in the protest peak network. We identified four main topics across all tweets: 1) protest campaigns and information, 2) fundraising activities for BLM by fans of a K-Pop group, 3) anti-protest messages and conspiracy theories, and 4) spam and irrelevant messages with the hashtag #BLM. These topics illustrate that discussion about BLM during the George Floyd protest is highly polarised,

with significant engagement from both supporters and opponents of the movement. We then summarised the distribution of topics and automated users across the four community clusters to address RQ5.1: What types of online networks emerge within the discourse surrounding the Black Lives Matter protests (see Data and Methods, topic modelling, community detection).

Our findings indicate that users operate in a highly polarised, hybrid political communication network during the protest peak. Table 5.1 shows the descriptive statistics and distributions of automated and human users for all communities during the protest peak. Four large community clusters with distinct structures, aims, goals, and sentiment levels have emerged, each with varying levels of bot engagement.

**Table 5.1:** Summary of network metrics and characteristics for each community cluster.

| **Metrics** | **Cluster 1 ('Battlefield')** | **Cluster 2 ('K-Pop fans')** | **Cluster 3 ('Anti-protest')** | **Cluster 4 ('Pro-protest and disruption')** |
|---|---|---|---|---|
| **Number of vertices** | 40,606 | 27,004 | 8,495 | 2,146 |
| **Number of edges** | 109,714 | 116,058 | 21,415 | 5,872 |
| **% of bots** | 52.89% | 71.13% | 45.23% | 59.92% |
| **% of humans** | 31.65% | 15.69% | 37.52% | 22.18% |
| **Average density** (human subgraph) | 0.033 | 0.1077 | 0.004 | 0.005 |
| **Average clustering** (humans subgraph) | 0.0187 | 0.1835 | 0.0579 | 0.026 |
| **Average reciprocity** (humans subgraph) | 0.0044 | 0.002 | 0.004 | 0.0001 |
| **Dominating topic** | 'Pro-protest' (46.5%); 'Anti-protest' (41.1%) | 'K-Pop fans fundraising' (84.9%) | 'Anti-protest' (46.9%) | 'Pro-protest' (53.1%); 'Spams' (29.20%) |

Community cluster 1 ('The Battlefield') is the largest cluster, comprising 40,606 users engaged in debates about the protest. Approximately 46.5% of the edges in this cluster are pro-protest messages, while 41.1% are anti-protest messages. This content has led to a highly fragmented community structure (0.033 density, 0.012 clustering). The community features a clear core-periphery arrangement, but due to the active communication, it has the highest percentage (0.44%) of mutual ties. This aligns with existing research on protest communication networks.

Notably, about half of the users discussing pro- and anti-BLM messages are social bots, highlighting their significant presence in both protest and anti-protest information diffusion.

Cluster 2 ('K-pop Fans') is a dense network with many mutual ties(0.107 density and 0.183 average clustering) composed of human K-Pop fans (15.69%) and pro-protest users exhibiting automated-like behaviour (71.13%). These accounts, claiming to be fans of the K-Pop group BTS, formed a close-knit network by actively reposting each other to spread their message and mobilise support for Black Lives Matter, matching BTS's $1 million donation through their fundraising campaign.

Clusters 3 ('Anti-Protest') and 4 ('Protest Campaigns and Spam Disruption') both exhibit extreme, negative attitudes towards BLM. These communities are fragmented, with a dense core broadcasting messages and a peripheral structure spreading them further. In Cluster 3, about 45% of users are automated, mainly disseminating anti-protest messages, including conspiracy theories about BLM's aims and motivations. However, the cluster has the second highest percentage of mutual ties (0.4%). Cluster 4 presents a more complex scenario, with edges predominantly spreading protest-related information. However, 29.2% of the edges are spam messages using the #BLM hashtag, such as interracial pornography, likely a strategy by automated accounts to flood timelines, distract users from protest information, and disrupt information diffusion and protest mobilisation.

### 5.2.2 Bot impact on human connection

To address RQ5.2 and RQ5.3, this study examines the prolonged effects of interactions with different social bots on human-human connections. Previous research suggests that bots can have a lasting impact on political discourse networks and significantly influence their structure. Therefore, this study aims to investigate whether interactions with bots affect human connections within these networks and the persistence of such interactions.

Our findings indicate that interactions with social bots during protest peaks lead to a significant decline in pre-existing human-human connections. It is essential

to measure this impact at both the individual and community levels to capture the full extent of the changes. At the individual level, we used changes in ego network clustering as a proxy for human network cohesiveness. This metric, which reflects how closely connected an individual's immediate network is, has been widely adopted in prior research to evaluate the influence of external factors on personal network cohesion. At the community level, we applied the Louvain community detection method to identify community clusters and measured the decrease in the percentage of human-human edges relative to the total number of edges in the network. Previous studies have used the proportion of edges between specific types of nodes relative to all edges as a measure of connectivity within a network. Following this approach, our metric assesses the overall connectivity between humans within communities by capturing the proportion of direct human interactions in the broader network.

To estimate the impact of human-bot interactions, we analysed the differences in these metrics before and after the protest peak. Overall, social bots significantly decrease the cohesiveness of human networks, as indicated by the decline in these metrics.

Figure 5.1 visually illustrates this process using individual level networks. While community-level analysis is also essential, the individual-level visualisations offer a clearer and more illustrative minimalist example of the observed phenomena.

In Figure 5.1, red dots represent users, blue dots represent bots, and the connections in the network are color-coded (human-human in red, bot-bot in blue, and human-bot in purple). Panels A-C and D-F demonstrate changes in ego network clustering for users with varying levels of bot interaction with ego networks(low: 2 interactions, high: 6 interactions) during the peak. The illustrations provide an example that interactions with social bots lead to a marked decline in pre-existing human-human connections in cases of high bot interaction. Despite starting with similar levels of human-human connections, networks with high bot interaction show a significant reduction in these connections post-peak,

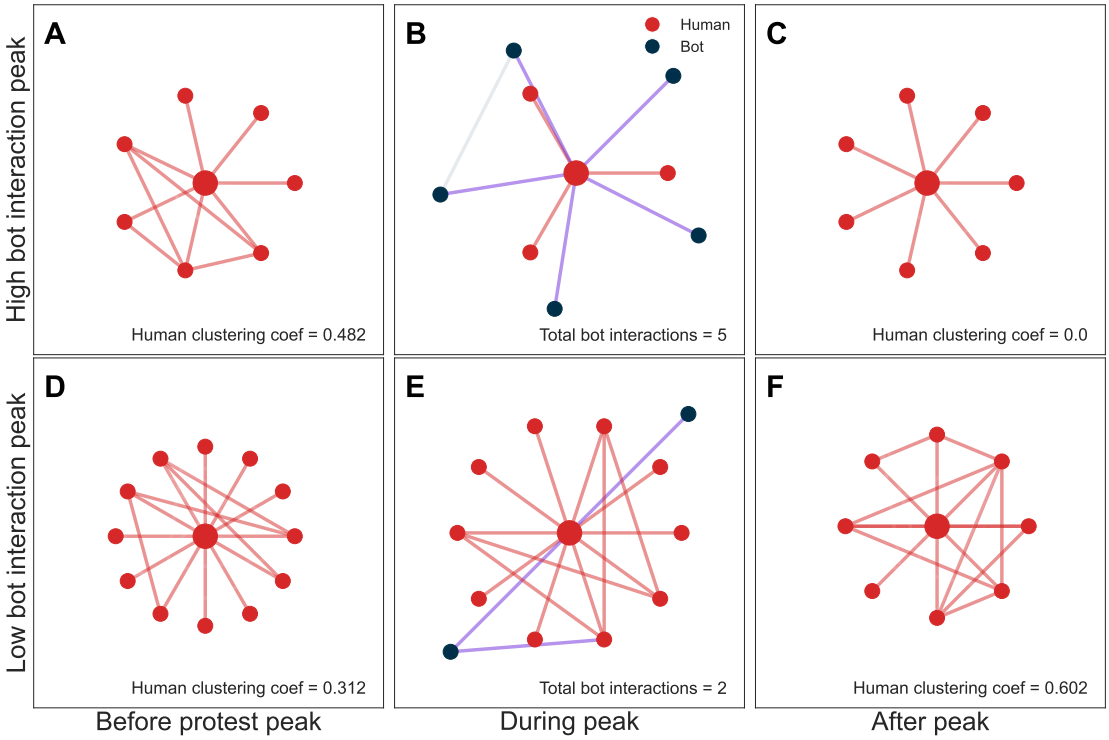



**Figure 5.1:** Two example of changes in ego network clustering before and after protest peak for users with high bot interactions (Panel A-C) and low bot interactions (Panel D-F). Red nodes represent human users, and blue nodes represent bots, with ego (large red node) in the centre. Changes in clustering coefficients are calculated based on subgraphs that contain only human nodes and edges, excluding self-loops. All edges are unweighted.

while those with low bot interaction maintain their connectivity and continue to facilitate active discussions.

We conducted OLS regressions to statistically confirm the effect, and to determine if the effects of different types of bot interactions vary. Figure 5.2 A and B illustrate the predicted changes (point estimators) in ego network clustering and community-level connectivity, with blue point estimators representing the predicted changes, differentiated by the number of human-bot edges in ego networks and communities.

Our analysis confirms that bot interactions lead to an overall decline in human-human connections at both the individual ($p < 0.05$) and community levels ($p < 0.001$). However, the effects differ based on the type of bot involved. At the individual level, we found that astroturfing bots primarily drive these changes, with a highly significant effect ($p < 0.001$), while interactions with non-astroturfing bots

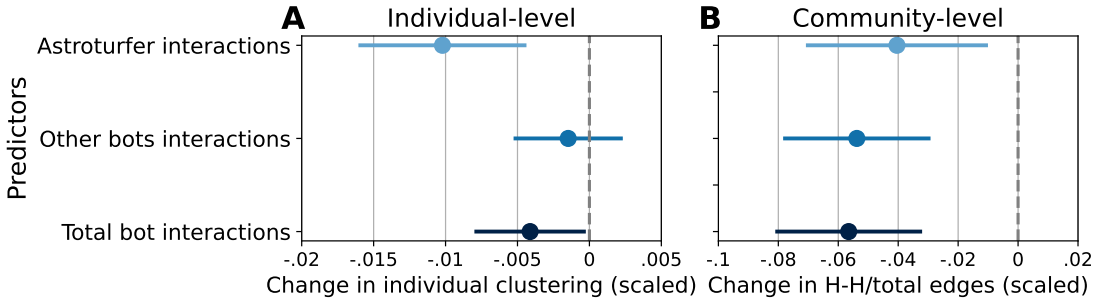



**Figure 5.2:** Point estimators of bot interaction's impact to changes in individual and community level connections. Panel A: predicted changes of human connection at the individual level, measured by clustering coefficients of the ego network. Panel B: predicted changes of human connection at community level, measured by changes in the proportion of human-human edges among all edges in the sub-community. Panels separate predictions (by colour) for astroturf bots (light blue), other bots (medium blue) and all bots (dark blue). The dependent variables are scaled to normalised distributions ranging from -1 to 1.

did not produce a similar impact ($p = 0.488$). At the community level, interactions with both types of bots resulted in a decline in human-human connections, but the effect was stronger for non-astroturfing bots ($p < 0.001$) compared to astroturfing bots ($p < 0.01$). These conclusions are supported by linear regression models, which highlight the differing impacts of bot types on network cohesiveness at both levels.

These effects were estimated while controlling for network structural trait metrics that influence the clustering and amount of human-human connections, individual traits affecting online connections and behaviour in online activism, and original opinions towards the protest. Further robustness checks suggest that the effects hold true across social bot CAP thresholds ranging from 0.5 to 0.75, and with different thresholds for determining astroturfing bots. (See SI Table B.4, B.5, B.8-B.20 for detailed regression models and robustness checks.)

Figures 5.1 and 5.2 collectively suggest that social bots may be 'replacing' pre-existing human-human connections during the peak of the protest. During this time, social bots appear to coordinate their efforts to increase their visibility within the communication network. This could explain the differing effects of various types of bots at the individual level.

At the individual level, the decline in human-human connections is primarily driven by interactions with astroturfing bots rather than non-astroturfing bots. Previous research indicates that astroturfing bots are particularly effective at flooding users' timelines with a high volume of messages or strategically targeting susceptible users and high-profile individuals to maximize their exposure. This effect is more pronounced at the individual level because astroturfing bots directly engage with users, drawing their attention and encouraging them to interact with the bot-generated content. Interaction with those bots can overwhelm communication channels and reduce the quality of interactions among users. This can lead to a decrease in engagement and a weakening of human-human connections as users become frustrated or disengaged. As a result, during the protest peak, human-bot ties are formed and persist even after the peak has passed, effectively replacing pre-existing human connections. This shift disrupts the natural structure of human online social networks, leading to a measurable decline in human-human connections, as indicated by decreases in clustering coefficients, and an increase in human-bot interactions. The increased interaction with astroturfing bots at the individual level thus contributes to the observed breakdown of human network cohesiveness.

At the community level, the effect of astroturfing bots on the decline of human-human connections is also evident, but less pronounced. A likely explanation is that, unlike at the individual level, other types of bots, such as fake followers, may also contribute significantly to this decline in community-level connections. For example, the second highest percentage of bots in our analysis is comprised of fake follower bots. These bots artificially inflate the follower counts of certain accounts, creating a misleading sense of popularity that can divert attention away from genuine human interactions. This shift in focus may cause users to engage more with bot-amplified content rather than maintaining connections with other community members. The effect is more prominent at the community level because fake followers influence broader visibility and engagement patterns, affecting multiple users simultaneously.

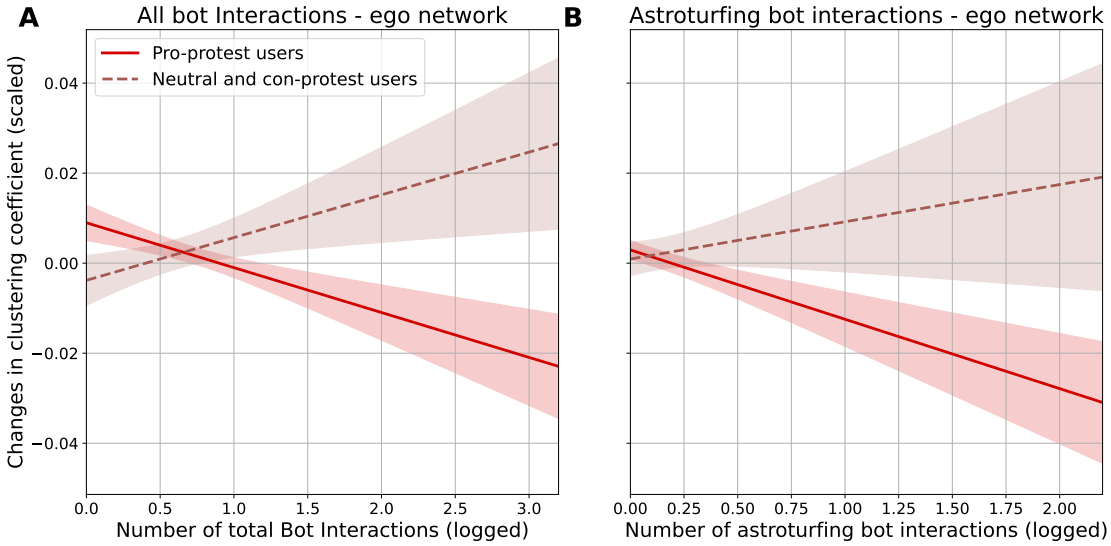



**Figure 5.3:** Line estimators of bot interaction's impact on changes of human connections at individual level. Panel A: predicted changes of human connection based on the number of total bot interactions. Panel B: predicted changes of human connection based on the number of astroturf bot interactions. Panels separate user groups by color for pro-protest users (dark red) and netural or con-protest users (light red). Panels show confidence intervals of the prediction around lines.

### 5.2.3 Moderation effects of pre-existing opinion

We also examined how pre-existing opinions (RQ5.4) of individuals moderated the impact of bots on human networks. Specifically, we were interested in whether the connections of individuals with different levels of support for the protest (pro, con, and neutral) were affected differently by bots, and whether individuals in different positions within the online network experienced varying levels of impact.

Regarding opinions, past research has suggested that previously held opinions are important in determining whether users are influenced by social bots. Pre-existing opinions affect how people react to bots, and bots may strategically target users with specific opinions to maximise their communication effectiveness. To explore this, we assigned each user an opinion category (pro-protest, neutral, or con-protest) based on their social media posts during the protest. The classification was conducted using computer-assisted pattern confirmation with ChatGPT (see Data and Methods and SI section Opinion categorisation using ChatGPT). We then compared changes in the clustering coefficient (at the individual level) and

the percentage of human connections (at the community level) among users with different protest support levels.

Our findings indicate that the decline in connections due to bot interaction primarily occurred among pro-protest users at the individual level. Figure 5.3 illustrates the predicted effects of overall (A) and astroturfing (B) bot interactions on human connections at the individual level. For both types of bot interactions, pro-protest users (represented by green lines) showed a predicted decrease in the clustering coefficient of their human-human ego networks as the number of bot interactions increased during the protest peak. In contrast, neutral and con-protest users (represented by navy lines) exhibited an increase in network cohesiveness with more bot interactions. The effect was statistically significant for both overall and astroturfing bot interactions (interaction term $p < 0.001$).

We observe no similar effect at the community level, likely because larger networks are more resilient, maintaining cohesion through indirect connections despite individual fragmentation. While bots disrupt personal connections, broader communities remain intact through shared goals, narratives, and interactions like hashtags or retweets. Moreover, individual disruptions may not accumulate enough to destabilise the overall community structure. (See SI Model Tables B.6-B.7 for full regression tables.)

## 5.3 Discussions

We found that social bots in Black Lives Matter (BLM) discussions formed distinct network structures based on varying motivations, similar to previously observed botnets. Bots coordinated to spread anti-BLM narratives, partisan content, conspiracy theories, and spam to disrupt mobilisation, while pro-protest activists and neutral entities, such as K-pop fans and news media, used bots to amplify protest news and fundraising efforts. In this polarised and fragmented network, anti-protest astroturfing bots is associated with a decline in human communication, affecting both individual and community-level interactions. Their impact varied based

on users' pre-existing opinions, with pro-protest users experiencing the strongest effects. In this section, we will discuss the implications of these findings.

Social bots have long been prevalent in online social networks, particularly in political communication. Previous research has shown that social bots work together online to replicate or spread specific messages, such as during Brexit[13], the US presidential election[143], or the Yellow Vest protests[23]. Past work suggested these bots form distinct network structures, like dense 'botnets.'[13]

In our study, we identified similar structures formed by social bots in online social networks, clustered by highly polarised opinions and motivations to participate in protests. In the context of Black Lives Matter (BLM) discussions, bots coordinated to spread anti-BLM, partisan, and pro-Trump messages. They also astroturfed conspiracy theories and other spam messages aimed at disrupting protest mobilisation among activists. However, pro-protest activists and neutral entities, such as K-pop fans and news media, also used automated accounts to disseminate protest news and fundraising campaigns. Our community detection analysis showed that automated accounts are active across various communities in BLM communication, spreading diverse messages within the discourse.

This is consistent with existing research[13, 23] regarding the prevalence of bots during online political events, where large-scale automated activity is observed. However, what stands out in our findings is that bot users remain active even after the protest has subsided. This persistence is notable because it suggests that these bots are not just deployed for short-term disruptions but are designed to maintain influence and continue engaging within the network over time. This long-term activity likely reflects a more sophisticated strategy aimed at sustaining narratives, influencing ongoing conversations, or keeping certain issues alive well beyond the protest' s peak, which is a notable deviation from the expected decline in bot activity post-event.

Based on bot community detection, we present the first evidence quantifying the causal impact of social bots on human-human connections in activism on social media, persisting for months after the peak of protests. Our empirical evidence

shows that social bots, especially con-protest astroturfers, cause a decline in human communications in online social networks. Interaction with social bots is linked to a decrease in human-human clustering in BLM social networks at both individual and community levels. This effect is particularly evident with interactions involving astroturfing bots compared to other types of bots. This represents a significant shift in the dynamics of online social networks, indicating that astroturfers not only manipulate information flow during critical events like protests but also lead to long-term changes in network structures.

These findings suggests that automated agents, whose behaviours differ markedly from those of humans, impact human-human network structures in a different way than human users online did, thereby affecting cohesion and connections of human networks. This finding indicates that social bots represent a novel factor in digital interaction dynamics that should be studied as a distinct entity by scholars in organisational studies, sociology, and social network analysis. The underlying mechanisms driving this phenomenon remain unclear, which urges further research to clarify the processes involved.

The empirical evidence from our study also reveals that activists and pro-protest audiences are particularly vulnerable to the influence of social bots, extending the impact of these bots beyond traditional political events like presidential elections[89, 75, 100]. This finding has significant implications for understanding how bots can affect civil political participation, suggesting that their disruptive potential is not limited to formal political processes but also extends to grassroots movements and social activism. The fact that both pro-protest and con-protest bots target pro-protest users highlights a critical vulnerability within activist networks, where the intended support from pro-protest bots might unintentionally monopolise discourse, while con-protest bots strategically introduce divisive content. This chaotic information environment can weaken the connection of activist communities, reducing their effectiveness and cohesion.

The broader implication of these findings is that social movements, which rely heavily on online platforms for organisation and communication[5], may be partic-

ularly susceptible to bot interference. The bot-induced decline in human-human connection within protest networks (social cohesion) is alarming as it is essential for protest mobilisation. Empirical evidence has demonstrated the importance of those ties to protest information diffusion and protest participation[305, 135]. Furthermore, cohesion is linked to a common political identity, social relations within a community, and orientation towards the common good[306]. These features are critical for mobilisation and can influence the outcomes of protest movements[147]. Therefore, a decline in cohesion poses a significant threat to civil society, given the essential role of protests in modern political processes and democracy.

Our study reveals that bots do not affect all users equally; they tend to target specific groups, particularly activists and pro-protest users. To build on these findings, future research should explore the specific strategies bots use to target different user groups and how these strategies influence the broader protest dynamics. Understanding these targeting tactics will be essential for developing targeted interventions to safeguard the integrity of online political discourse and enhance the resilience of social movements against digital manipulation.

Additionally, future studies should investigate the types of community structures that may offer protection against bot influence. For instance, networks with strong reciprocal ties[307] or a higher number of triangles (indicating tightly-knit groups) may be more resilient to bot-driven disruptions. Protesters are often connected through such structures[136], which may build their resilience against bot interference. Identifying and reinforcing these protective structures could make online communities less vulnerable to the adverse effects of social bots. Furthermore, this research could inform the development of 'good bots' that activists could use to strengthen pro-protest networks and support their causes, further enhancing the resilience of social movements.

Our research was conducted prior to the development of advanced AI models such as ChatGPT[218] (launched on November 30, 2022). Large Language Models (LLMs) facilitate the activities of social bots, making them more capable of mimicking human interactions and initiating connections within communication

networks[93]. Given our findings that social bots contribute to the decline in human connections, the potential dangers posed by these advanced bots are even more pressing. Consequently, further research is needed to understand why this is happening and to develop strategies for regulation and mitigation. Such research would be valuable for social media platform service providers, activists, and other stakeholders, including legislators.

In summary, our study highlights the urgent need for comprehensive research to address the disruptive effects of social bots on online social networks, especially regarding political activism. Understanding and mitigating these effects are important to maintaining the integrity of digital communication networks and supporting democratic processes.

## 5.4 Data and methods

### 5.4.1 Data

This research examines the human and bot networks related to the Black Lives Matter (BLM) protests on X (formerly Twitter). We focused on discourse around the protests following George Floyd' s death on 25 May, 2020, and identified 6 June to 8 June, 2020, as the peak period of protest activity based on existing descriptive statistics.

To capture relevant discourse, we collected all tweets posted during this peak period containing keywords such as 'BLM,' 'George Floyd,' and various related terms with slight spelling variations. This data was gathered using the Twitter Academic Product track API, which provides a comprehensive archive of historical tweets. In addition to tweet content, we collected metadata including user IDs. This allowed us to retrieve additional user information via the Twitter user lookup API. Our final dataset includes 2.5 million tweets from more than one million users between 6 June, 2020, 00:00 and 8 June, 2020, 00:00.

(See SI data collection; see Table B.1 for the full list of keywords.)

### 5.4.2 Bot detection

We adopted and fine-tuned a bot detection method which involves cross-validation and solid robustness checks developed by our past research[293]. This involved using both the publicly available tool, Botometer[35], and a self-trained bot detection algorithm. 'Social bots' were defined as users with a Botometer score threshold above 0.65 and identified as bots by our algorithm. Users with scores ranging from 0.5 to 0.65 were labeled as 'unknown', and the rest were classified as 'humans'. (See SI bot detection for the full procedure).

Past research indicates that the concept of social bots is nuanced, requiring further refinement. Botometer provides a typology of social bots, categorising them into 'astroturfers', 'spammers', 'financial bots', and others. As suggested by Li et al.[293], astroturfers are particularly prominent in political discussions and likely to have a significant impact. Therefore, we further categorised bots into 'astroturfers' and 'other bots' for detailed analysis.

Astroturfers are social bots that coordinate to spread messages on social media, often by posting the same message and retweeting each other in a short period. To classify them, various Botometer scores were considered: astroturfers, fake followers, spammers, and financial bots. If a bot scored above 2.5 (above average) in the 'astroturfer' category and highest in this category compared to others, it was classified as an 'astroturfer'. Other bots were classified as 'others'. Robustness checks with different cut-off thresholds for astroturfers ranging from 3 to 3.5 were also conducted, with no significant difference between main effects (see SI Model Tables B.17-B.20 for details).

To ensure these astroturfing bots are indeed bots, the total CAP (Complete Automation Probability) was considered. Previous work have shown that bot probability significantly affects bot behaviour and traits. Thus, classified astroturfers were further restricted to bots with a CAP higher than the median of all bots. Overall, 13,810 users were classified as 'astroturfers', 37,172 as 'other bots', 13,875 as 'unknown', and 27,221 as 'humans'.

### 5.4.3 Network construction and filtering

We began by constructing a bi-directional, unweighted retweet network (RT network) where users are represented as nodes, and retweets or quotes between users serve as edges. In this network, all edges indicate the mere presence of a connection between users. For instance, multiple interactions between a pair of nodes are aggregated into a single edge to simplify the analysis. This approach helps to understand if interactions with bots impact connections between human users. This is an approach that was adopted by past research on online communication network[308, 309], and particularly in the online political communication context[310].

In this context, 'retweets' refer to posts shared by hitting the retweet button without additional comments, while 'quotes' may include extra commentary. Each retweet or quote was recorded as a directional edge, with the original tweet poster as the starting node and the retweeter or commenter as the end node. Additionally, self-loops were recorded for users who retweeted or quoted their own tweets or posted original tweets. Self-loops were included as they affect online discourse and protest mobilization, according to past studies. Users without any original tweets or retweets were excluded from the RT network.

We then focused on a subset of users who actively interacted with bots during a 48-hour snapshot of the BLM communication network. Given that online communication networks in protests often exhibit a hierarchical core-periphery structure[140], filtering the network to its core structure is essential to examine social cohesiveness and impact, as suggested by past work[311, 312]. We employed the k-core metric for network filtering, which identifies the maximal subgraph where each node is connected to at least k other nodes[313]. Users with a k-core value greater than two were included, resulting in a network of 92,078 users and 372,932 edges. This filtering was performed using Python's NetworkX package[314].

To assess the impact of bot interactions on network cohesiveness among human users, we created two additional network snapshots: one before and one after the BLM protests. The 'before' network included data from 1 May, 2020, to 2 May, 2020, while the 'after' network encompassed data from 1 September, 2020, to 2

September, 2020. Directed communication networks were constructed based on retweeting information. We collected full timeline data for these periods, instead of filtering the edges by BLM-related keywords

### 5.4.4 Opinion categorisation using ChatGPT

To achieve the research goals, users' opinions about Black Lives Matter or civil rights movements in general were categorized. Previously-held opinions significantly impact users' reactions to their interactions with bots[58]. Users' stances on protests were measured using a score from -1 to 1: -1 indicates strong opposition, 1 indicates strong support, and 0 indicates neutrality or unrelated discussions. For further analysis, these scores were grouped into three categories: scores from -1 to -0.1 as 'Anti-BLM (Con protest),' scores from 0.1 to 1 as 'Supporters (Pro protest),' and scores from -0.1 to 0.1 as 'Neutral.'

User opinions were classified based on tweets from the dataset, examining all interactions (human replies to bots) among sample users and bots. Each user's opinion was assessed by categorising opinions in all tweets and retweets before and after bot interactions.

To classify users' opinions, OpenAI' s large language model, ChatGPT 3.5[218], was used, a method supported by several studies[219, 220]. For each user, the model was provided with instructions to classify their opinion on climate change protests based on their tweets. The model then produced the aforementioned scores. These scores were generated by analysing users' entire timelines in the network snapshots before bot interactions (pre-interaction opinion), after bot interactions (post-interaction opinion), and human replies to bots during interactions (opinion during interaction). (See SI Opinion categorisation for detailed information on prompt engineering and verification.)

### 5.4.5 Topic modelling

To understand the context in which human-human edges and human-bot edges formed in, we adopted bi-term topic modelling to understand the themes that

emerged in BLM-related discussion during the protest peak. This was done with the bi-term topic models specifically designed for modelling short text[228, 229] that learn topics by modelling word-word co-occurrence patterns[230].

Overall, we found 15 topics in our BLM discourse corpus during protest peak, and further used both computational and manual methods (based on thematic similarities) to group the 15 topics into 4 broader topics. As the protest peak at the George Floyd protest is made up of skirmishes of protests across the United States[46], the topics identified are rather fragmented and highly centred around specific small-scale protest events. We therefore further clustered them into broader topics through qualitative interpretation of the topic keywords, so that the new topics better capture the opinion, information diffusion and organisation in the communication networks. We then found four broader topics across all protest-related discourse during the protest peak.

(See more details about topic modelling procedure and details of topic models in SI Topic Modelling. See Figure B.1 and B.2 for topic model evaluation metrics, Table B.2 for the topics identified.).

### 5.4.6 Community detection

We conducted both individual and community-level analysis to illustrate changes in social cohesiveness induced by social bots from another perspective. For community level analysis, we used sub-communities formed during the protest peak as the smaller unit for our analysis, a procedure taken by past research[122]. To achieve this, we first used community detection methods to detect 156 communities available for analysis. Community detection involved a two-stage process conducted on the communication network during the peak of the protest for both research and practical purposes.

In the first stage, we employed the label propagation community method to partition the full graph into 8 large communities, in which four of them formed meaningful clusters (see Table 5.1). This method assigns labels to nodes in a network and propagates these labels based on local connectivity patterns. We chose

this method due to the relatively large size of our network (12,918 nodes) and its known efficiency and scalability. Label propagation methods identify communities based on diffusion, a characteristic relevant to traits such as ideologies, social identity, and social bots, which strongly impact clustering.

In the second stage, we further divided each large community into smaller, fine-grained sub-communities using Louvain community detection methods. Louvain community detection is an algorithm designed to identify clusters or communities within networks by optimising the modularity measure, which evaluates the strength of division of a network into communities. It effectively reveals the underlying structure by maximising the density of connections within communities compared to connections between communities[315].

This method is optimal for our analysis because it efficiently handles large networks and provides a clear delineation of community structures[315], making it suitable for examining fine-grained interactions within the network. Overall, we identified 167 sub-communities for statistical analysis. We retained only those sub-communities containing at least two human nodes, ensuring that each community could form meaningful human-human interactions. This selection process resulted in 156 communities for further analysis.

We then created three snapshots of communities using the same set of nodes detected above: before, during, and after the protest peak. By observing the evolution of the same community during and after the protest peak, we could understand how human network cohesion changed at the community level.

### 5.4.7 Statistical analysis design

We employed a dual-level approach to assess the effects of bot interactions on network cohesiveness in the context of online activism. This approach includes both individual-level and community-level analyses. Methodologically, this dual approach is necessary due to differences in sampling and the unique insights that each level provides. Theoretically, analysis at both levels allows us to comprehensively assess the effects of bot interactions on cohesiveness, providing stronger

empirical foundations for theories of network studies and social movements if results are consistent across both levels.

Sampling differences between these two levels also make separate analyses necessary. Individual-level analysis focused on users who were active before, during and after the protest peaks. Community-level analysis considers the collective interactions and connections within broader social structures, and therefore includes users who were not active before or after protest networks. By analysing both levels, we can more accurately capture the dynamics of bot influence on online protest networks. This comprehensive approach ensures that we address potential biases and variations that could arise from focusing solely on one level of analysis.

### 5.4.8 Individual-level analysis

We estimated the effect of bot interaction on network cohesiveness at the individual node level. To achieve this, we selected a sample of users active in communication networks before, during, and after the protest, with a k-core value greater than two. We eliminated all isolated nodes from the pure human k2 network during the protest peak, as these nodes had no human-human edges initially and would not appear in the network without interaction with bots.

Human-human connection (individual-level cohesiveness) is assessed by examining the change in local clustering coefficients during and after the peak of the protest for each individual in the network. The clustering coefficient is a measure of the extent to which nodes in a graph tend to cluster together[316]. Specifically, a local clustering coefficient measures the proportion of a node's neighbors that are connected to each other. This metric is used in existing literature[317] as an indicator of network or structural cohesiveness. By analyzing its change before and after the peak of the protest, we can understand how cohesiveness among humans changes with the involvement of a protest event. The metric is typically calculated as the ratio of the number of edges between a node's neighbours to the maximum possible number of edges between them, as shown in the following formula[316]:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

We then measured 'bot interaction' as the number of unweighted degrees that go from humans to bots or vice versa in the protest peak network. In other words, it is estimated as the total number of unique bots with which humans interacted during the peak of the protest period. Two retweets from the same bot are counted as one interaction. This approach provides a clearer picture of how extensively a user is exposed to bot influence, as interacting with multiple unique bots likely reflects a broader and potentially more impactful exposure to automated behaviour. Furthermore, we estimated the total number of both astroturfing and other bot interactions. This is because of the significance of astroturfing bots in impacting human behaviour, as shown in both past work[75].

We used Ordinary Least Squares (OLS) regression to estimate the effect of bot interaction on human network cohesiveness. To achieve this, we incorporated a series of control variables. First, we control for network metrics that may impact clustering coefficients, including the number of degrees and the clustering coefficient for human network before protest peak. Furthermore, previous work has found that two user-level features strongly influence users' roles in protest-related information diffusion[113]: the ratio between friends and followers, and the ratio between messages received and sent. Therefore, these two user traits are included to estimate their effect on changes in network cohesiveness. Additionally, we controlled for other user-level features that are relevant to their cohesiveness in the community, such as account age, number of statuses, and number of favourites. Account age measures how long a user has been active in the online community and is associated with cohesion. Posting frequency (number of statuses) and liking activity (number of favourites) measure how active a user is in the online community, which are also relevant to the level of network cohesion[318].

To examine the moderation effect of pre-existing opinion, we also included opinion in part of our analysis to measure its moderation effects. Users' previously held opinions strongly impact their credibility and susceptibility to social bots[58],

and bots may target users with a specific kind of opinion[69]. The dominant type of user (71.9%) in our communication peak network is pro-protest. 21.4% of users were neutral and 6.5% held a negative opinion, and as the percentage of either is relatively low to form a meaningful sample, we decided to combine the two categories for further analysis. We then included a binary control variable indicating whether the user is pro-protest (labelled as 1) or not pro-protest (0). We analysed whether users with different opinions (positive, or not positive) experienced different changes in their connections with bot interactions. We conducted a series of robustness checks using subsets of data from users who held positive or non-positive opinions towards BLM during the peak of the protests.

### 5.4.9 Community-level analysis

We measured community-level network connection using the percentage of human-human edges relative to the total number of edges in the network. This method is particularly effective because it provides a direct and quantifiable measure of genuine human connectivity within a hybrid network that includes both bots and humans. By focusing on the proportion of human-human edges, we can assess how much of the network's interaction remains between humans, which is important for understanding the impact of bots on natural human connections. Previous studies have demonstrated that edge-based metrics are reliable for evaluating network connectivity[319], as they allow researchers to track shifts in interaction patterns within specific groups, such as humans, relative to the entire network structure. Using the ratio of the percentage of human-human edges after the protest peak to that before the peak, we can precisely measure changes in human connectivity likely caused by social bots, providing a clear indicator of the extent to which bots have disrupted or reshaped the network.

It is important to note that the percentage of human-human edges is a proportion, meaning it naturally decreases as the proportion of bot interactions increases. This introduces an inherent limitation, as the presence of bot nodes has a scaling effect on the estimated impact of bot interactions. As a result, the actual effect

of bot interactions may be slightly smaller than estimated. Future research could improve robustness by incorporating absolute measures, such as controlling for the total number of bots, to ensure that estimates are not influenced by proportional scaling. In future work, such approaches could provide a more precise assessment of bot influence in online networks.

At community-level, human-bot interaction was also quantified as the number of edges within a community connecting a human node and a bot node. We also differentiated between astroturf-human edges and other bots-human edges to measure the distinct influence of the astroturfing of bots on human communities. OLS regression was then employed to assess the effect of such interaction on human-human connection.

In addition to the key explanatory variable of human-bot interaction, we controlled for several factors. Network metrics included the overall scale of the network (total nodes and edges) and the percentage of human-human connections within each community before the protest peak. Individual-level controls encompassed the ratio of friends to followers, the ratio of messages received to sent, account age, and the number of statuses. We also examined moderation effects by including pre-existing user opinions. For individual-level features, we employed the average values of individual features for all human users in each community as community-level controls.

# 6

# The strategies of social bots in shaping online activism discourse

## Contents


**6.1 Introduction . . . . . 135**
**6.2 Theoretical framework and research gaps . . . . . 139**
6.2.1 A refined discursive opportunity structure framework . . 139
6.2.2 Bots and visibility . . . . . 141
6.2.3 Bots and resonance . . . . . 145
6.2.4 Bots and legitimacy . . . . . 148
**6.3 Methodology . . . . . 150**
6.3.1 Case and data . . . . . 150
6.3.2 Computer-assisted content analysis . . . . . 151
**6.4 Findings . . . . . 153**
6.4.1 Visibility . . . . . 154
6.4.2 Resonance . . . . . 160
6.4.3 Legitimacy . . . . . 167
6.4.4 Overview of social bots' role in the two protest cases . 174
**6.5 Discussions and conclusions . . . . . 177**

# Preamble

This chapter addresses RQ3: What strategies do social bots use to influence discourse in online activism? Specifically, it applies a computer-assisted content analysis to examine bot strategies that impact the discursive opportunity structure (DOS) of activists in the two protest cases: Extinction Rebellion and Black Lives Matter.

This chapter builds on the causal claims established in Chapters 4 and 5 about how bots influence human behaviour and connections. It then shifts focus to examine what kinds of bot activities shaped human actions and relationships. It addresses the research gap identified in Chapter 2, particularly in the areas of 'protest and social media–mobilisation tactics' and 'social bots–behaviour and strategies.' This chapter contributes to RQ3 by systematically conceptualising bot strategies and their impact on online activism discourse, offering vivid, detailed examples of these strategies. It enhances both empirical and theoretical understandings of social bot activities and advances the theory of online activism by refining the existing framework. See Appendix C for supplementary information.

**Publication status:** This chapter is being prepared for submission to *New Media & Society*.

## 6.1 Introduction

In the last decade, online activism has been heavily shaped and facilitated by grassroots political discussion online[320, 321]. Activists rely on social media platforms such as X (formerly Twitter), Facebook or Instagram to carry out online activism[322]. During the process, discourse, together with images and video, has become a major tool how activists promote social changes. With their messages online, protesters diffuse information, provoke emotional reactions and achieve support from their audience[28]. By targeting the general public, activists aim

to amplify the reach and impact of their messages[112], transform engaged individuals into potential activists[323], and exert pressure on decision-makers through widespread public sentiment[324]. Consequently, much of their efforts are focused on making use of online discourse to effectively engage with the general public.

Because of the importance of discourse for protests, past research[325, 326] has proposed a theoretical framework called the 'discursive opportunity structure' (DOS) to analyse it. The framework refers to the conditions under which discourse can successfully diffuse and resonate with the public and lead to social change. It was originally developed in the context of traditional media, where DOS highlights how activists navigate through the available discursive opportunities to diffuse their messages. Before social media, DOS was largely shaped by media gatekeepers who controlled the visibility and framing of activist messages[326]. However, the rise of social media has transformed the DOS, allowing activists and other stakeholders to independently impact the discursive opportunity of their messages[327].

A significant factor that may influence the modern DOS is the presence of social bots, or automated accounts, across various social media platforms. These bots are prevalent in online spaces and have been observed in numerous protest movements, such as the climate change protests[43, 293] or the Black Lives Matter movement[296]. Automation allows social bots to generate vast amounts of content at minimal cost and rapidly engage with human users on social media platforms[12]. This capability gives them significant influence over the diffusion of information related to specific agendas online, thereby actively shaping the DOS of activists.

Despite the importance of social bots in shaping the DOS, there is limited understanding of how they influence the discursive opportunities available to activists. Indirect evidence suggests that bots compete with activists to spread different messages[128], influencing public interactions[328] and support for protests[23]. However, we lack work that explicitly discuss how social bots impact political discourse. Therefore, it is important to investigate the specific strategies employed by social bots and their effects on the discursive landscape within protest movements.

Furthermore, investigating the discursive dynamics between bots and humans provides an opportunity to refine and update the original DOS theory within the context of social media. The DOS theory was initially developed during the era of traditional media, primarily conceptualised discourse as a means for activists to disseminate information[326]. However, recent theoretical developments and empirical studies indicate that discourse now functions as a form of activism itself[329, 330]. Social media introduces new dynamics, particularly through micro-level interactions between diverse agents such as social bots, activists and the general public. Traditional DOS analyses often overlook these interactions, yet they are essential for understanding the evolving nature of online protest discourse. There have been a few pieces of empirical work discussing the role of discourse in online activism[327, 331, 332], but we still lack comprehensive conceptualisation. By examining how social bots influence these dynamics, we can enhance the DOS framework to better align with the realities of modern online activism.

This research, therefore, aims to investigate how social bots influence the DOS of activists on social media during protests. Drawing on the framework by Koopmans et al.[326], which outlines three dimensions of DOS—visibility, resonance, and legitimacy—we will refine and adapt these concepts to the specific context of social media. Our analysis will explore how each dimension is shaped by the interactions between bots, humans, and their collective influence within the hybrid media environment of online protests.

To address this research topic, we analyse social media data on bot activities during two key protest cases: Extinction Rebellion (XR) and Black Lives Matter (BLM). XR is a global environmental movement that began in the UK in October 2018[44], while BLM focuses on raising awareness about police violence against African Americans and combating racism and inequality[191].

These cases are particularly suitable for our study due to their representativeness and potential for comparative analysis. Both involve large-scale protests that rely heavily on social media for mobilisation and information dissemination[39, 45], representing significant movements within liberal democratic contexts and focusing

**Table 6.1:** DOS definition and sub-dimensions

| Aspect | Definition | Sub-dimensions |
|---|---|---|
| Visibility | The extent to which the general audience is exposed to activists' messages | Global exposure<br>Influence capacity |
| Resonance | The ability of protest frames to provoke reactions from other actors in the discursive field | Emotional resonance<br>Cultural resonance |
| Legitimacy | the level of support for a social movement, as expressed through communication and online activism | Legitimacy of protest frame<br>Legitimacy of activists' image<br>Extent of participation in protest |

on major public concerns: environmental sustainability and racial equality. Preliminary observations suggest that social media bots have a discernible impact on the discursive structures during these protests[296, 293]. which highlights the relevance of the DOS in these movements. This provides us with ample data to conduct a detailed analysis of how bots influence the DOS of activists in these contexts. Additionally, we can explore whether bot strategies and human interaction patterns are consistent across both cases, potentially offering insights into the generalisation of social bot operations in online activism within liberal democracies.

The next section of this paper begins by presenting the theoretical framework that underpins our research questions, focusing on the three dimensions of the DOS in online activism and how social bots may impact them. We then refine this framework to better fit the context of online activism, review existing empirical research on bots according to these dimensions, and identify research gaps. With extensive computer-assisted content analysis of bot activities and human reactions in the two cases, we then identify bot strategies. More specifically, we examine how existing strategies manifest in these cases and whether new strategies emerged that were not captured in previous research. We then estimated the extent to which pro- and con-protest bots applied these strategies using a sample of bot posts. Finally, we discussed the implications of these findings in the broader context of digital activism and online discourse.

## 6.2 Theoretical framework and research gaps

### 6.2.1 A refined discursive opportunity structure framework

This research investigates the impact of social bots on online activism by shaping the discursive opportunity structure (DOS) for activists. Discursive opportunities were initially defined as the factors in the public sphere that allow certain messages to diffuse. Since the public sphere has limited space for messages, different stakeholder groups must compete for message dissemination. Thus, the DOS framework was introduced to analyse factors influencing this competition[326]. In the context of activism, the primary stakeholders are the groups of activists who organise to promote their protest goals. Furthermore, the DOS framework was initially introduced in the context of traditional media. In the context of traditional media, media outlets act as gatekeepers, determining which messages are disseminated into the public sphere. Therefore, the dynamics of DOS primarily focus on how activists' messages disseminate and elicit reactions through traditional media.

Koopmans et al.[326] originally suggested that DOS has three aspects: visibility, resonance, and legitimacy. Visibility refers to how noticeable certain messages are to the public, which determines their dissemination. Resonance indicates the extent to which a message provokes public reactions. Legitimacy describes the degree to which an action gains community support, affecting message diffusion and the perception of the protest and its agents. Consequently, the analysis of DOS focuses on how activists interact with traditional media to achieve visibility, resonance, and legitimacy. However, social media introduces new dynamics for civic political participation, requiring adaptations to the framework.

Social media brings about two major changes that are not addressed in the current DOS framework. First, the existing DOS framework did not incorporate the ability of discourse to trigger sociopolitical change. Instead, it views discourse solely as a means to diffuse information about other collective actions like protests. Recent theoretical developments and empirical studies acknowledge that discourse

itself serves as a form of activism[329, 330, 109, 333]. Discourse can shape individual opinions and actions at a micro level[334], form 'connective action' organisations online at a meso-level[324], and ultimately trigger sociopolitical change with or without collective offline action[109], as seen in cases such as the #MeToo movement[335]. A more refined DOS framework should incorporate discourse's ability to catalyse social change, rather than focusing solely on the its effect on information diffusion.

Second, the traditional DOS framework focuses on a single central traditional media agent, like TV or newspapers, disseminating information in a one-way flow to a passive audience. In contrast, social media creates a dynamic, decentralised system where various agents—media, activists, the public, and social bots—interact in a complex, multi-directional communication environment[5]. Social bots, in particular, interact with human users in rapid and unique ways, leading to varied reactions that collectively shape the audience' s perceptions of protest frames[12]. Past DOS studies, however, have neglected the presence of social bots in online discursive structure. Analysing the mechanisms of interaction between bots and audiences is essential to understand how these interactions influence the DOS of activists. Due to the distinct nature of social bots, such as their ability to disseminate numerous messages in a short period, new sub-dimensions of DOS may emerge that were previously unrecognised in human and social media settings.

Therefore, we need to broaden and adapt the DOS framework proposed by Koopmans et al.[326] to account for the complexities of social media. This revised framework will serve as the basis for outlining our research questions and guiding our analysis. In building our framework for analysis and research questions, the concepts of all three dimensions—visibility, resonance, and legitimacy—are refined and developed according to relevant literature and the specific context of social media. During the process, we review past work and discuss their theoretical definitions and empirical measurements adopted in the social media context. This serves two purposes: first, we use our refined concept and empirical measurements together to define sub-dimensions of the concept. The concepts are very broad, and

as we aim to use them to illustrate bot activity, we need a systematic and more detailed framework of DOS. Second, they provide empirical evidence of whether bots impact DOS based on those measurements.

Finally, we consider existing literature on potential mechanisms of how bot activities and interaction with the audience impact DOS of activists. Although the authors of these past works did not originally organise their findings about social bots in this manner, nor did they explicitly mention them, we will reorganise their findings using this framework to identify potential research gaps.

See Table 6.1 for a summary of a refined DOS framework in the hybrid social media context, including refined definitions of each DOS dimension and potential sub-dimensions. It is worth noting that new sub-dimensions may arise in empirical study in the Findings section.

### 6.2.2 Bots and visibility

#### Dimensions

The first aspect of DOS is visibility. Visibility is originally defined as the extent to which the general audience is exposed to activists' messages. Definition of this aspect has been relatively stable, from when Koopman et al.[326] first raised the concept, to when scholars discuss information diffusion in the context of social media. It deals with the extent to which the general audience is exposed to activists' messages. Visibility has been recognised as the most significant and apparent aspect of discourse opportunity because it determines the diffusion and potential impact of the activist's messages[327].

While we find the original definition of visibility by Koopman et al.[326] useful, we propose incorporating two sub-dimensions to better capture its complexity in the context of social media. In the original DOS framework, visibility was largely controlled by media gatekeepers. However, in the context of social media, activists and other stakeholders have the ability to increase the visibility of a message independently. Based on both theoretical and empirical past work[122, 336], we

argue that there are two potential sub-dimensions or measurements of visibility in this context: global exposure and reach, and influence capacity.

The first dimension is global exposure and reach, indicating how many unique users are exposed to the content. Actors on social media compete for visibility within the same public sphere, as the total public space for a message to diffuse online is limited. As such, if a certain message is seen by more users or reaches more audience, it becomes more visible in the public sphere, and the rest of message become less visible. In past empirical work, visibility metrics on social media platforms was also commonly developed to measure this mechanism. The most naive way to measure it is the total number of reposts or total number of views, which is an effective indicator of how far the message has diffused, and how many users are exposed to the messages[337, 338]. A more refined way of measuring exposure is to estimate the global reach of the message - for example, the ratio of number of reposts divided by the total number of reposts during the given time period[279, 280]. Some empirical studies[122] also used the pagerank value according to percentage reach and the number of followers in X as a measurement of global exposure.

The second dimension of visibility is influence capacity. While global exposure measures the reach and visibility of a message, influence capacity focuses on the ability of an agent (such as an activist or a bot) to distribute and amplify content within a social media network. This includes the agent's centrality in the information network and their potential to shape discourse by facilitating the spread of information across a wide range of users.

Agents on social media seek to amplify their messages to increase the likelihood of them being seen and shared by others, especially in the context of online activism[112]. However, the total amount of a specific message online does not necessarily guarantee its influence. The capacity of an agent to be at the centre of communication networks and drive the conversation is important for creating impact[339]. This sub-dimension captures an agent's ability to strategically amplify

their messages through their positioning within the network and their engagement with others, rather than the emotional reaction it may provoke in the audience.

As its definition centres around agents, most empirical measurements of influence capacity centres around agents on public sphere. Past work had multiple approaches to develop metrics for influence capacity. They first adopted a social network perspective, and used centrality of a node in communication network to measure influence capacity[122]. This is because nodes with a higher centrality is more connected or influential within the network. Furthermore, Castro-Gonzalez et al.[340] measure influence of a agent through the ability of interactions to extend the reach of content (amplification bandwidth, the ratio of interactions to the number of followers), as well as the effectiveness of converting followers into interactions (funnel Bandwidth, the ratio of interactions (such as likes, comments, and shares) to total followers).

**Bot strategies**

Although no research explicitly mentions how social bots impact the visibility of messages, previous studies indeed highlighted the potential of social bots to impact message visibility. Social bot operators exploit bots to automatically disseminate large volumes of messages rapidly. Two behaviours that could affect activists' visibility have been identified, and they contribute to the two dimensions of visibility accordingly. While they also influence other aspects of the DOS, their primary impact is on visibility, and therefore we elaborate on them in this section.

The first strategy mentioned by past studies that bots have used is astroturfing. Keller et al.[75] describe astroturfing as a strategy where agents simultaneously post similar messages to generate information cascades. It is intrinsic to the nature of social bots and platform mechanisms, such as automation—the ability to post a large number of messages in a short period of time. Therefore, this behaviour has been observed during time-sensitive political events like the Brexit Referendum and the US presidential election[13, 42]. Bots often follow and retweet each other

to reinforce astroturfing, contributing to trending political hashtags and drawing human users into these conversations[94, 95].

Astroturfing contributes to the first element of visibility, known as global exposure. Global exposure measures how broadly a message spreads and how many unique users encounter it. To achieve high global exposure, bots must amplify the volume of messages and ensure their wide dissemination across the network. Astroturfing is particularly effective for this purpose because it allows bots to generate a high volume of content in a short period, making use the platform' s algorithms to artificially boost the visibility of certain messages. By reposting and retweeting content in a coordinated manner, these bots create the illusion of widespread support or interest, which can deceive users into believing they are witnessing a significant, organic movement[75]. This manipulation increases the likelihood that real users will engage with the content, further enhancing its reach and reinforcing the bot-driven narrative.

Social bots increase message exposure by connecting with non-automated users, which is the second strategy identified by past work. They initiate connections and communication at a high frequency and strategically select targets to maximise message impact. For instance, during the Catalan independence referendum, social bots connected with influential accounts to spread their messages[18]. They also used violent content to flood user timelines and increase exposure to controversial material. In cases like the US's withdrawal from the Paris Agreement[42] and the 2016 presidential election[100], bots targeted individuals susceptible to disinformation or aligned with their partisan views.

Such connection and targeting contributes to the second aspect of visibility - influence capacity. As mentioned, influence capacity assesses how well an agent can enhance an actor' s ability to be at the centre of the network, and influence users. To improve influence capacity, social bots make use of platform mechanisms to maximise their impact on online communication. Targeting specific users is a particularly effective strategy for leveraging platform mechanisms. Platform

algorithms tend to prioritise frequently posted content and interactions with influential accounts[341]. Additionally, these algorithms recommend content based on users' interests and preferences[342]. As a result, bot-generated messages can appear more prominently in the feeds of susceptible users, especially when the bots directly target them.

Through those two strategies, social bots can either hinder or promote the visibility of certain agendas in the public sphere, as suggested by past research. These two strategies allow them to impact the two dimensions of DOS, thereby attracting collective attention, promoting user engagement, and ultimately increasing the visibility of messages. Through the process, If the agenda is pro-protest, it enhances the visibility of activists' messages. If the agenda opposes the protest, it competes with and diminishes the visibility of activists' messages.

However, despite previous research outlining bot strategies, we still lack insight into the specific content of the messages they amplify. While it is well-established that bots engage in astroturfing and target individuals, the nature of these messages —whether they promote pro-protest or con-protest visibility—remains unexplored. Additionally, although past work has described bot strategies like astroturfing and targeting, there is little understanding of how these strategies manifest in the context of protests. We need to investigate specific cases of astroturfing and targeting within protests, determine the type of content being spread, identify whether the messages are pro- or con-protest, and assess the scale of each side's activity.

Therefore, we raise the following research question (RQ6.1):

> How do social bots use astroturfing or user targeting to promote or undermine visibility?

### 6.2.3 Bots and resonance

The second type of discursive opportunity recognised is resonance. Previous work, beginning with Koopmans et al.[326], defined 'resonance' as the ability of a message to provoke reactions from other actors in the discursive field. They identified two types of reactions: consonance, when the reaction is supportive, and dissonance,

when the reaction is not supportive. Social media has created new discursive opportunities for activists to achieve resonance with their protest frames, but it has also increased competition due to the sheer volume of messages. Theoretical and empirical work in social movement studies[343, 344] generally adopts similar definitions, defining resonance as the ability of a protest frame or anti-protest frame to provoke reactions.

While the original definition offers a useful starting point, we believe it is necessary to introduce additional sub-dimensions of resonance to more accurately measure the role of discourse in online activism. Resonance, as a critical element of protest actions, has various functions such as attracting audience attention on key issues, fostering solidarity among activists and supporters, and catalysing collective action[345]. Given its centrality to effective protest, the original definition' s general approach to resonance does not fully capture these fine-grained aspects. By incorporating specific sub-dimensions, we aim to provide a more comprehensive understanding of how resonance operates within the context of online activism and protest.

There are two kinds of resonance most relevant to protest: cultural[346] and emotional resonance[347], which we identify as sub-dimensions of DOS resonance. Cultural resonance is described as 'an appeal on the basis of common experience'[346]. It involves the extent to which a message aligns with and evokes a response based on shared values and identities, motivating people to react and further diffuse the message. If the resonance is beneficial for people to support the protest frame, it creates consonance and an advantage for activists in the DOS. If it aims to be anti-protest, it creates dissonance and a disadvantage for activists in the DOS.

Emotional resonance, on the other hand, does not rely on shared experiences. It engages personal feelings such as anger, joy, sadness, or shame relevant to the protest. To achieve this, messages often use radical or extreme expressions to target individuals' emotions[348]. Similar to cultural resonance, emotional resonance can

either support or hinder a protest frame depending on the emotions elicited and the target of those emotions.

Despite the equal importance of both emotional and cultural resonance to activists, past research on social media and bots has predominantly focused on emotional resonance, largely neglecting empirical research of cultural resonance across various dimensions. Nonetheless, we will summarise existing work on emotional resonance and bots here.

Empirical studies have found that bots have successfully provoked reactions and altered human sentiment (emotional resonance), particularly in social movements, towards the negative side. They have been shown to amplify emotions, transferring anger to humans and exacerbating online conflicts[18, 349]. This was observed in both non-protest cases, such as COVID-19 discussions[350, 349], and protest cases like climate change protests. Li et al.[293] further confirmed that direct interactions with bots led to more negative sentiments in discussions about protests. Both micro-level evidence and macro-level descriptions indicate that bots provoke emotional reactions in human users.

Two potential strategies may have been adopted by social bots to achieve such negative emotional resonance, as suggested by past work. First, social bots often post and promote messages aimed at evoking strong human emotions[18]. This is particularly prevalent in discussions about political events, such as the Catalan Referendum, the Brexit Referendum, and Japan's general election, as well as public events like COVID-19 discussions[18, 13, 97, 349]. In these contexts, social bots orchestrate efforts to disseminate inflammatory, highly controversial, negative, and hyperpartisan messages designed to inflame strong human emotions like anger and fear. Second, social bots amplify existing messages posted by human users[24], particularly those from political actors. For instance, during Russian interference in the 2016 U.S. presidential election, social bots targeted conservative or pro-Trump users[102], and magnified the content and sentiment of these messages when reposting them. Furthermore, in the case of COVID-19

conspiracies[350], bots would retweet neutral human-posted messages along with negative or pessimistic comments.

There are several gaps in current research on bots and resonance that we aim to address. First, previous studies have not examined cultural resonance strategies of social bots, so we need to determine whether bots engage in this and identify the strategies they use, with specific examples. Second, although two key bot strategies of emotional resonance have been identified, they have not been analysed in the context of protests. Since these strategies are not protest-specific, we need to investigate how they manifest in protest settings. Third, we must explore whether new strategies have emerged that were not previously documented. Finally, it is not clear whether social bots aim to create resonance or dissonance within the context of online activism. This involves identifying how many bots are designed to generate pro-protest resonance versus those that create anti-protest resonance. Addressing these gaps will require a systematic exploration of bot activities in the context of DOS resonance. Hence, we raise the following research questions (RQ6.2):

> RQ6.2: How do social bots use strategies to enhance or reduce emotional or cultural resonance?

### 6.2.4 Bots and legitimacy

The third aspect of DOS is legitimacy, which refers to the level of support for activists' claims as reflected in the reactions of third parties in the public sphere. Other research in online activism[351, 352, 353] further define legitimacy as the perception that a movement' s organisers, goals, tactics, and overall existence are appropriate, justifiable, and acceptable within the broader social and political environment, as perceived by third-party agents such as the general public, media, or neutral stakeholders.

Past definitions of legitimacy were developed within the context of traditional media environments and offline protests[353]. Given the evolving dynamics of online activism, it is essential to refine the definition of legitimacy in the context of digital discourse and its impact on protest movements. Studies of protest legitimacy

in the context of social media recognise that it serves as a discourse field providing new means and contexts for protesters to gain legitimacy for themselves and their activities[354, 355]. As social media increasingly becomes a platform for activism and political participation, activists use discourse to induce social change directly, rather than merely diffusing messages about other forms of activism, such as offline protests. Thus, legitimacy should also encompass the extent to which third parties participate in or claim to engage in civic political participation.

Therefore, we expand the definition of DOS legitimacy to 'the level of support for a social movement, as expressed through both communication and participation in online activism'. Building on past work, we propose three sub-dimensions of legitimacy. Since our new definition of DOS legitimacy includes protest participation, the first sub-dimension is the extent of participation in protests.

The other two sub-dimensions are derived from previous empirical studies in political science and psychology, which quantify perceived legitimacy through survey questions, or qualitatively investigated the dimensions of legitimacy. Both surveys and interviews suggest perceived protest legitimacy concerns two other categories: the goals and means of the protest, and the movement organisers[352, 356, 357]. Thus, we argue that DOS legitimacy should include two additional sub-dimensions: first, the perceived validity and justifiability of the protest's aims, goals, and actions; second, the legitimacy of activists, or the credibility and appropriateness of the individuals or groups conducting the activity.

There are a limited number of research studies on whether bots have changed users' perceived legitimacy of a protest, but they do not separate between the sub-dimensions: legitimacy of activists and legitimacy of protest frames. Lab-based experiments suggest that bot interactions can change opinions on protest means and goals[291], and computer simulations indicate that around 2% to 4% bot participation in a network may be sufficient to shift the opinion climate[129]. Li et al. found that increased interactions with social bots are associated with decreased support for pro-protest users in the case of Black Lives Matter (refer to Chapter 5 of this paper). This evidence suggests that bots can influence support

levels for protests, impacting legitimacy and shaping the DOS for activists. There are currently no empirical evidence on whether bots impact the actual participation in online activism or not.

There is a lack of literature specifically addressing the strategies that bots use to alter the perceived legitimacy of a protest. However, related research offers some insights. Social bots are known to spread misinformation and conspiracy theories, particularly during the COVID-19 pandemic and the U.S. midterm election[100, 101]. If such misinformation targets protests, it could lead to skepticism regarding the aims, goals, and organisers of the protest, thereby diminishing the second and third dimensions (perception of protest frames and activists) of legitimacy in the DOS. On the contrary, pro-protest bots, such as 'woke chatbots'[16] may exist to help increase civic participation. This could potentially promote actual participation in a protest, and impact the first dimension of legitimacy (protest participation). However, no existing research has formally investigated their impact.

Therefore, we identified two research gaps. First, although there is empirical evidence implying that bots may alter protest legitimacy, the specific strategies they use remain unclear. Additionally, there is a lack of understanding regarding the extent to which bots affect legitimacy and whether their influence is positive or negative. Addressing these gaps requires a detailed investigation into the strategies employed by bots and their overall impact on protest legitimacy. To address those two gaps, we raise RQ6.3:

> RQ6.3: What strategies do social bots use to influence the perceived legitimacy of protests, focusing on public perception, the image of activists, and levels of participation?

## 6.3 Methodology

### 6.3.1 Case and data

As mentioned, this research adopts a computer-assisted content analysis (CACA) approach to conduct an in-depth content analysis of two cases of online activism:

Extinction Rebellion (XR) and Black Lives Matter (BLM). Our aim is to investigate how social bots and human users' reactions to bots interact to impact the discourse opportunity structure (DOS) of human users, using the analytical framework detailed above.

We filtered a dataset of bot posting records and user discussions about social bots to conduct further analysis. The dataset was sampled from two larger datasets of social media posts on X (formerly Twitter) during the peak of the two protests. For both cases, we used the Twitter Academic product track API[213] (available by then) to collect all tweets containing protest-related keywords during the peak protest periods. Additionally, as this research focuses on analysing bot activities, we also gathered user activities and applied a self-developed bot detection method[293] to categorise all users as either 'bots' (automated accounts), 'humans,' or 'unknown' (see SI Case and Data and Bot Detection).

To conduct an in-depth analysis and address our research questions, we further filtered the corpus using a series of procedures, resulting in six distinct datasets. These datasets consist of social media posts from X during the peak of these protests, including interactions between humans and bots, as well as human discussions about social bots. Table 6.2 shows all the datasets used for our analysis. Quotes from the datasets will be referenced in the format of 'dataset + #post index.' For example, the first post from the XR1 dataset will be referred to as 'XR1 #1.'

### 6.3.2 Computer-assisted content analysis

We employ a hybrid, computer-assisted content analysis (CACA) approach to address our research questions. Content Analysis (CA) is a qualitative research method used to systematically interpret and categorise textual data based on themes and patterns identified through human coding and analysis[358]. CACA extends this methodology by integrating computational tools, such as natural language processing (NLP) algorithms, with traditional qualitative techniques to efficiently analyse large textual datasets[359].

**Table 6.2:** Datasets used for computer-assisted content analysis.

| Dataset | Source | Content | Timeframe | Sample Size |
|---|---|---|---|---|
| XR1 | X ('Twitter') | Bot retweets during the Extinction Rebellion (XR) protest | 14/11/2019-05/12/2019 | 2,513 |
| XR2 | X | Bot original posts during the Extinction Rebellion (XR) protest | 14/11/2019-05/12/2019 | 567 |
| BLM1 | X | Bot retweets during the Black Lives Matter (BLM) protest | 06/05/2020-08/05/2020 | 5,383 |
| BLM2 | X | Bot original posts during the Black Lives Matter (BLM) protest | 06/05/2020-08/05/2020 | 1,438 |
| USR1 | X | User posts mentioning social bots during the BLM and XR protests | 14/11/2019-05/12/2019; 06/05/2020-08/05/2020 | 437 |
| RD1 | Reddit | Posts and responses regarding bot activities in online political activism | Since 2020 | 62 |

Note: All search keywords included their various forms, including plural/singular iterations, diverse spellings, and other variants.

CACA is particularly suitable for analysing large amount of textual data, and given our sample size, effective for investigating how bots influence the DOS of human users through social media interactions. Following existing recommended practices of CACA[359], our research procedure involves initial computer-assisted data filtering, manual pattern detection, and automated pattern confirmation.

The datasets used in this analysis include four sub-datasets of bot activity, one sub-dataset of human perception of social bots (USR1), and one sub-dataset of social bot discussions on X (RD1). To understand how bot activities impact the three dimensions of DOS—visibility, resonance, and legitimacy—we examined the most retweeted 500 bot-generated messages and all human discussions about social bots. We developed a codebook based on close reading and theoretical frameworks, and each tweet was coded according to this codebook.

We conducted manual coding for USR1 and RD1, and used computational tools to assist in coding the larger bot activity datasets. Specifically, we employed OpenAI' s large language model, ChatGPT 3.5[218], to classify bot-generated messages

based on predetermined categories, with the model outputting the corresponding code. Following this, we performed descriptive statistics on the coded data to quantify the prevalence of specific bot strategies. This statistical analysis helps identify patterns and trends within the data, providing a clearer understanding of how different bot activities influence visibility, resonance, and legitimacy in online discourse.

We used Taguette[360] for developing the codebook, coding, and data analysis. Ethics approval for the study was obtained from the Central University Research Ethics Committee at the University of Oxford (CUREC ref: SSH_OII_CIA_21_109.).

(Refer to SI computer-assisted content analysis for method selection rationale and further details on the procedure.)

## 6.4 Findings

The interaction between bots and humans significantly shapes the discursive structure of activists on social media during protests. In the XR and BLM cases, about 40% of users exhibited bot-like behaviour, highlighting the substantial influence bots have on the DOS of activists by engaging with human users online. This influence occurs through three key dimensions of DOS: visibility, resonance, and legitimacy. Table 6.3 summarises these findings.

By 'interaction between bots and humans,' we refer to the activities carried out by automated accounts and the responses of human users to these activities. Bots are designed to fulfil specific purposes, such as spreading large volumes of messages, while human users, who are neither bots nor activists, engage with this content, amplifying its impact.

Social bot strategies are categorised as either destructive or constructive based on their effects on DOS. Destructive strategies disadvantage activists by reducing visibility, creating negative resonance, or undermining support for protest frames. Conversely, constructive strategies benefit activists by increasing message visibility, evoking positive emotions, and garnering support for protest agendas.

**Table 6.3:** Bot strategies grouped by DOS dimensions

| Dimensions | Subdimensions | Destructive strategies | Constructive strategies |
|---|---|---|---|
| Visibility | Global exposure | Con-protest astroturfing | Pro-protest astroturfing |
| | Influence capacity | Targeted engagement with celebrities | Targeted engagement with protest keywords |
| Resonance | Cultural resonance | Appeal through common values | Appeal through common values |
| | | Fabricate human identity | Promote trustworthiness of bot |
| | Emotional resonance | Anti-protest direct provocation and amplified neutrality | Pro-protest direct provocation and amplified neutrality |
| Legitimacy | Protest participation | Hinder protest engagement | Promote protest engagement |
| | Legitimacy of protest frame | Criticise protest frame | Justify protest frame |
| | Legitimacy of activists | Portray activists negatively | Portray activists positively |

In the following sections, we will explore how social bots shape each dimension of DOS. For each dimension, we will outline the sub-dimensions and the bot strategies that influence them, providing detailed descriptions and representative examples. We will also assess the scale and usage of these strategies through computer-assisted categorisation and descriptive statistics. During the process, we consider the context of platform mechanisms and bot-human interactions rather than analysing bot activity in isolation.

### 6.4.1 Visibility

Social bots make use of two major strategies—astroturfing and targeted engagement—to influence the visibility of messages within online protest movements. These strategies are employed by both pro- and con-protest bots, each impacting different aspects of visibility in the digital public sphere. As past research has identified these strategies, our focus is on describing the context in which those strategies are adopted, and providing a detailed statistical description of the scale of their

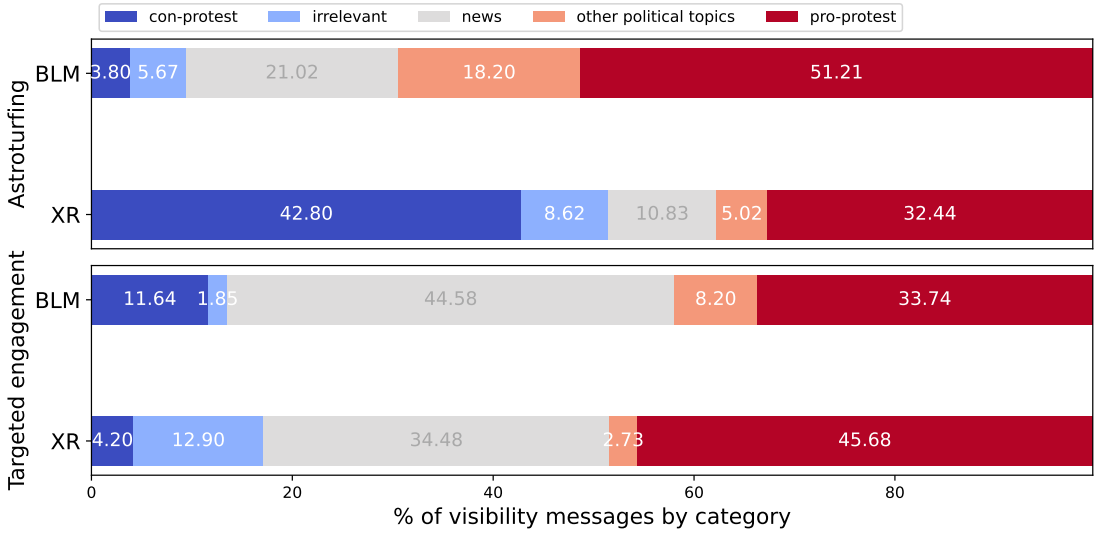



**Figure 6.1:** A breakdown of bot-generated messages by content, divided into two types of bot activities: astroturfing (top) and targeted engagement (bottom). This analysis covers both BLM (Black Lives Matter) and XR (Extinction Rebellion) cases. All values represent the percentage of messages within each category, relative to the total number of messages for that specific activity (astroturfing or targeted engagement) and case (BLM or XR).

usage. We will conduct an in-depth analysis of the content posted by these bots, detailing the proportions of various message types associated with each strategy.

### Astroturfing

Social bots first frequently employ astroturfing to amplify the visibility of messages during the protests. As discussed in 'Theoretical framework and research gaps', astroturfing contributes to the first dimension of visibility: global exposure. This strategy, which involves coordinating mass reposts and retweets, significantly boosts the exposure and perceived popularity of certain messages, manipulating social media algorithms to dominate online discussions. Astroturfing has been extensively examined in past research, and we found similar patterns in our dataset, where bots used this tactic to create the illusion of widespread support or interest.

We observed significant usage of astroturfing behaviours by social bots that affected activists' visibility. There were 5,383 unique instances of astroturfing retweets for Black Lives Matter (BLM) and 2,513 unique instances for Extinction Rebellion (XR). In total, this amounted to 55,674 retweets for BLM, representing

25% of all messages, and 94,043 retweets for XR, accounting for 30% of total messages. With our dataset of bot tweeting records and a count of total number of bot retweets, we find a significant amount of messages were reposted simultaneously by social bots, who were probably orchestrating to promote the global exposure of the tweet on users' timeline. Hereby 'astroturfing' cases are defined as bot-generated messages with an above-median number of suspected automated retweets (9 for BLM, and 4 for XR). For the recorded astroturfing cases, each message was retweeted on average 10.5 times by social bots for BLM and 4.2 times for XR cases.

Furthermore, we find that social bots used astroturfing strategies to both promote and hinder the visibility of protest agendas on social media in our cases. The behavioural patterns of astroturfing were similar for both constructive and destructive bots, but there were notable differences in the content they promoted and across the two protest cases. Figure 6.1 shows the statistics of all types of content astroturfed across both protests.

Pro-protest bots accounted for 67.7% of the astroturfing records across the two protests (71.9% for BLM and 42.4% for XR). The higher percentage of pro-protest activity in the BLM case suggests a more concerted effort by bots to amplify supportive narratives, while the lower percentage of pro-protest bot activity in the XR case may indicate a more fragmented, polarised bot activity landscape. Further content analysis reveals that these bots in BLM and XR primarily focused on spreading two types of content.

The first type of content that pro-protest bots astroturfed in the cases was 'pro-protest messages,' where bots promoted protests or mobilised campaigns, such as fundraising efforts. For example, many highly retweeted messages by automated accounts in the BLM case were fundraising campaigns by K-pop fans supporting Black Lives Matter, aligning with their K-pop group's $1 million donation through the campaign. These messages accounted for a significant portion: 51.1% of all astroturfing records for BLM, and 32.34% for XR, where bots spread pro-protest messages and protest information.

The second type of content they astroturfed consisted of news or media reports of the protests, which could be neutral or pro-protest. Despite the neutral tone, these messages generally spread information about the protests and their progress, contributing to visibility, and can therefore be considered a form of constructive bot activity. Overall, news reports made up a notable proportion of astroturfing in both protest cases (20.9% and 10.7% for BLM and XR, respectively). These messages likely originated from media bots and automatic accounts that post news or from activism bots that retweet updates on protest progress. At the same time, con-protest bots constituted of 32.3% of all astroturfing records across the two protests (31.1% of BLM, and 58.4% of XR, respectively). They roughly spread three kinds of content.

Destructive bots primarily spread 'anti-protest messages,' aiming to promote anti-protest sentiment or mobilise counter-protest movements, such as the #All-LivesMatter campaign. This is the focus of astroturfing bots for XR: 42.7% of destructive social bots engaged in astroturfing to disseminate messages that directly criticised the goals, methods, or participants of the protest. For example, a widely retweeted astroturfing tweet in this category stated, 'The irony of these idiotic teens chanting "fuck the police" whilst on a climate change protest when only an hour later the very same police force were facing a potential suicide bomber on #LondonBridge, saving many lives in the process of tackling him. #LoveOurPolice.' (XR1, #1) This tweet received over 30,000 retweets, illustrating the bots' success in amplifying anti-protest rhetoric. For BLM, the percentage was 8.7%, suggesting that con-protest bots did not focus on spreading this kind of message.

The other two categories of con-protest astroturfing messages involved content not directly related to protests but aimed at undermining the visibility of activist agendas. The second con-protest category, 'other political propaganda,' focused on promoting unrelated political agendas, such as partisan preferences. Typical astroturfing content in this category included attacks on political figures (often liberal or Democrat) or expressions of support for certain political parties or figures (often conservative or right-wing, such as pro-Trump messages). This type of

content was predominantly used by BLM astroturfers (18%), but less for XR astroturfers (8.52%). For example, one of those cases (BLM1, #3226) argued that 'Conservatives protested peacefully but were branded as racists' while accusing Liberals as 'marxists'. This strategy was disruptive as it diverted attention from the protest's core issues, diluting the message and potentially polarising the audience.

The third category of destructive astroturfing messages included content entirely irrelevant to the protests, such as pornography or advertisements tagged with protest related tags. These irrelevant messages may have served dual purposes: promoting specific agendas for financial or other gains and disrupting the spread of protest information. Although these irrelevant messages constituted a small proportion of all astroturfing records (5.67% for BLM and 8.62% for XR), they were frequently flagged by human users during discussions. For example, one user commented, 'WARNING !!! There is currently some sort of bot creating fake profiles tweeting porn with hashtag #ExtinctionRebellion...It seems to be a concerted attack happening right now.' (USR1, #11). Human users' vigilance to bot messages suggests that despite the low volume of those messages, they may reduce the visibility of critical protest information and mobilisation messages.

**Targeted engagement**

Another strategy used by social bots is 'targeted engagement'. Targeted engagement involves bots repeatedly posting the same message, often as replies or comments that target verified users or specific hashtags. As mentioned, it contributes to the second dimension of DOS by enhancing influence capacity, and therefore is also observed in our dataset. We have observed that bots could post the same message 30 to 50 times to interact with users central to the communication network.

Targeted engagement strategies employed by social bots significantly increase the visibility and impact of specific protest messages. In total, 1,348 original messages (1.6% of all BLM original bot messages) from the BLM dataset and 567 (4.5% of all XR original bot messages) from the XR dataset were posted three or more times, suggesting a deliberate strategy of targeted engagement and replication

to increase visibility. Despite their smaller proportion, they had a significant impact: BLM-targeted engagement messages received 82,948 retweets (16.9% of all retweets received by BLM bots), while XR-targeted engagement posts received 9,530 retweets (10.1% of all retweets received by XR bots). This shows these target engagement messages' considerable influence capacity, which significantly affected the visibility of the messages they aimed to promote or hinder.

Bots use targeted engagement to disseminate both pro-protest and con-protest messages, often spreading similar categories of content as those seen in astroturfing, though with slightly different proportions. Compared to astroturfing cases, targeted engagement tended to be more pro-protest in BLM and XR. Specifically, 78.1% of the BLM targeted engagement records were pro-protest, with the remaining percentage being con-protest (29.9%). Similarly, for XR, 78.9% of targeted engagement records were pro-protest, with 22.1% con-protest. Figure 6.1 provides a detailed breakdown of the percentages of different categories of pro- and con-protest messages across all targeted engagement cases.

Further in-depth analysis reveals that bots constructive or destructive to protest communication exhibited slightly different targeting preferences. Pro-protest bots focused on specific keywords like protest hashtags, such as a bot that commented on images found under the #blacklivesmatter tag, thereby increasing the visibility of their messages within relevant discussions. This strategy ensured that the bot' s message reached users who are actively engaging with protest-related content, amplifying the protest's message within its core audience. On the other hand, con-protest bots targeted verified users who often have large followings. By engaging with these high-profile accounts, the bots' messages reached a broader audience through the increased visibility of these influential figures. In both cases, bots significantly enhanced the influence capacity of activists, allowing them to shape public perception and influence online discourse by saturating the conversation with agenda-driven messages.

Overall, bots play a profound and significant role in influencing the visibility of messages in the digital public sphere, and we studied impacts that are both constructive and destructive across the two protest cases. Bots employed two main strategies—astroturfing and targeted engagement—that are particularly effective in dominating online discussions and enhancing message visibility. A significant proportion of bot activity was directed towards amplifying pro-protest messages. This finding contrasts with existing research that predominantly emphasises automated accounts spreading misinformation and disadvantaging activists' messages. Instead, our study reveals that a considerable portion (for example, 20% in all BLM bot retweets) of bot messages actually worked to increase activists' visibility, demonstrating the dual role bots can play in shaping online discourse.

Furthermore, when comparing the two cases, we uncover some different patterns of bot activities. The BLM case demonstrated a stronger emphasis on promoting pro-protest content through both astroturfing and targeted engagement, which reflects a more cohesive effort among bots to support activist messages. In contrast, the XR case exhibited a higher proportion of con-protest messages within its bot activity. This suggests a more complex dynamic in which bots not only amplified pro-protest narratives but also worked to undermine them.

### 6.4.2 Resonance

In our framework, we categorise the strategies used by social bots to achieve resonance into two types: cultural and emotional. Emotional resonance involves the bots crafting messages that elicit strong emotional responses, such as anger, fear, or empathy, to sway public opinion and increase engagement. Cultural resonance refers to the alignment of a bot's messages with the cultural, social, or identity markers that resonate with a particular audience. Bots achieve this by adopting identities, language, and symbols that are familiar and trusted by their target audience, thus making their messages more credible and impactful. Bots use a variety of narrative and behavioural strategies to achieve these and consequently shape DOS resonance, which is detailed and illustrated below.

The difference in the proportion of social bots aiming to achieve resonance between XR and BLM—49.6% for XR and 88.3% for BLM—can be attributed to the distinct goals and emotional appeal of each movement. BLM' s focus on racial justice and social equity may elicit more intense emotional and cultural responses, leading to a higher frequency of resonance strategies. In contrast, XR' s emphasis on climate change, while significant, might not engage the same level of emotional resonance, resulting in a lower proportion of bots using these strategies. This disparity highlights how the nature of the cause can influence the strategies employed by bots.

Table 6.4 shows the proportion of bot-generated messages in the aforementioned sample categorised into emotional resonance, cultural resonance, or no resonance for both constructive and destructive bots across the BLM and XR datasets. This table highlights how different bot types used these resonance strategies.

In the following sections, we will provide a detailed analysis of emotional and cultural resonance. For each category, we will include specific examples to illustrate how these strategies are employed by bots to either engage audiences or disrupt discussions, as well as the percentage of those categories across our sample. This analysis will offer insights into the varying approaches taken by constructive and destructive bots to influence public sentiment and achieve their respective goals.

**Emotional resonance**

We find that emotional resonance is a key strategy used by both constructive and destructive bots in the XR and BLM protest movements, but its application varies significantly between the two cases and between different bot stances.

For constructive bots, those linked to XR placed a strong emphasis on emotional resonance, which comprises a notable portion of their pro-protest activity. Emotional resonance accounted for 41.7% of all pro-protest bot activity in XR and 17.6% of all XR bot activity samples. BLM's constructive bots also engaged with emotional resonance, and it represented a slightly smaller share of their overall activity (36% of all pro-protest records in BLM and 24.4% of total BLM bot records).

**Table 6.4:** Distribution of Resonance Categories by Bot Type in BLM and XR Datasets

| Bot Type | Resonance Category | % of Same Bot Type (BLM) | % Overall (BLM) | % of Same Bot Type (XR) | % Overall (XR) |
|---|---|---|---|---|---|
| **Constructive** | Emotional Resonance | 36.0% | 24.4% | 41.7% | 17.6% |
| | Cultural Resonance | 62.2% | 42.1% | 29.6% | 12.5% |
| | No Resonance | 1.9% | 1.3% | 28.7% | 12.1% |
| **Destructive** | Emotional Resonance | 46.7% | 15.1% | 27.7% | 16.0% |
| | Cultural Resonance | 20.7% | 6.7% | 4.0% | 2.3% |
| | No Resonance | 32.2% | 10.4% | 66.6% | 38.5% |
| **Total** | | | 100% | | 100% |

The percentages represent the distribution of messages from each bot type across different categories in the BLM and XR samples. "Percentage of Same Bot Type" refers to the percentage of messages within a specific bot category (e.g., 36% of all messages from all constructive BLM bots demonstrate emotional resonance). "Overall" indicates the distribution of messages across all bot categories for each protest. The resonance categories are mutually exclusive, meaning each message is classified into only one type (emotional, cultural or no resonance), and the percentages sum to 100%.

On the destructive side, BLM bots were more effective at using emotional appeals to provoke strong reactions and disrupt discussions, while XR's destructive bots relied less on this strategy. Emotional resonance was more evident in BLM cases, where it represented 46.7% of con-protest activity and 15.1% of all sample activity. For XR, emotional resonance accounted for 27.7% of con-protest activity and 16.0% of all sample activities.

These statistics indicate that emotional resonance serves as a powerful tool for both types of bots, with its effectiveness varying between movements. The higher prevalence of emotional resonance among BLM's destructive bots highlights their ability to create immediate impacts, as emotional appeals can quickly polarise discussions and divert attention from core issues. Meanwhile, the significant presence of emotional resonance among both kinds of constructive bots points to a strategic approach to engage audiences and amplify their messages. This dual role of emotional resonance in shaping online discourse reflects the complexities of bot activity within these protest movements.

Social bots aim to achieve emotional resonance with protest frames by directly communicating with users through two distinct narrative strategies in their tweets.

Emotional resonance involves eliciting individual feelings such as anger, shame, resentment, pride, or excitement, or other emotions that are valuable in protest contexts. To achieve this, bots typically post messages designed to provoke strong emotional reactions. These provocative messages are crafted to elicit emotional responses from users, influencing opinions and driving engagement. This is achieved through two kinds of narrative strategies.

The first narrative strategy can be termed 'directive provocation.' Both pro-protest and anti-protest bots use explicit statements or actions designed to elicit an immediate emotional response from their audience. This approach aims to create emotional resonance that either benefits or undermines the protesters, depending on the bot' s objectives.

Constructive bots use directive provocation by engaging with the protest frames to evoke a wide range of emotions useful for protest. They often strive to elicit positive sentiments such as unity, pride, and confidence. For instance, one of the suspected automated users retweeted, 'Thank you so much Extinction Rebellion Australia for your OUTSTANDING work...wishing you a FANTASTIC GLOBAL CLIMATE STRIKE' (XR2, #86). By using words such as 'outstanding' and 'fantastic,' these comments generate positive reactions from a pro-protest third-party audience. Furthermore, wishing a 'fantastic global climate strike' provokes optimism and motivation, encouraging viewers to feel hopeful and inspired about the impact of the protest and the broader climate movement. However, the sentiment can also lean towards the negative, such as urgency, anxiety, or anger. One constructive automated-like account posted the message multiple times, 'People are suffering and dying from the climate emergency and it will continue to get worse,' to provoke a sense of urgency and anxiety in the audience, encouraging them to take action.

Destructive, con-protest bots achieve direct provocation by directly posting inflammatory comments or offensive remarks towards protest frames. For example, one tweet (BLM1, #5822) labelled BLM a 'Marxist puppet,' a term that can easily provoke negative sentiment by associating the movement with controversial

political ideologies for targeted audience (the U.S. general public). Additionally, bots graphically described riots and looting during BLM protests, which can incite fear and anger by highlighting disorder and criminal activity. These automated messages used provocative words such as 'disgusting,' 'sick,' and 'disgraceful' to intensify negative emotions.

Another narrative strategy for emotional resonance could be named 'amplifying neutrality'. More specifically, bots amplify opinions by commenting on protest-related news, so that they also elicit a emotional response from the general audience on social media. This strategy may be adopted to subtly influence public opinion without overtly taking a side, thus avoiding direct confrontation and maintaining a facade of neutrality. By engaging in this manner, bots can exploit emotional triggers and manipulate discussions, fostering engagement and spreading their agenda while appearing less biased.

Constructive bots achieve this by sharing protest-related news reports and tweaking the wording of the news titles to emphasise neutral statements in a positive light, thereby using context to evoke feelings of happiness, solidarity, or support. For example, a pro-protest account of XR activists with automated behaviour shared a news report titled 'Extinction Rebellion hunger strikers target UK political parties.' They paraphrased the activists' behaviour by describing it as 'hunger strikers demanding urgent action on the climate-eco crisis' and commented on it as 'a timely action' (XR1, #751). This conveys a sense of urgency and importance while framing the hunger strike in a positive light, thus fostering feelings of solidarity and support among the audience.

For destructive bots, amplification of neutrality usually involves highlighting a neutral comment or fact about protest in a provocative context. For instance, when The Guardian reported on XR activists' disruptive behaviour, such as gluing themselves to buses, the media used neutral wording. However, bots reposted these messages with comments like, 'always had a feeling XR was a tool made to assimilate would-be radical protesters and flush them down the drain' (XR1, #2158). This added judgement and satire, indirectly suggesting that XR activists

were unsafe and disruptive to public order, thereby provoking negative reactions towards the XR protest frame.

**Cultural resonance**

Cultural resonance is notably present in pro-protest bot activities, especially in the BLM context. Constructive bots in BLM prominently used cultural resonance, which accounted for 62.2% of all pro-protest activity in BLM and 42.1% across all BLM bot records. This high emphasis on cultural resonance is linked to BLM' s strong connection to racial identity and social justice, themes that align well with cultural narratives. Constructive bots in XR focused slightly less on cultural resonance, reflecting the movement's central concern with climate change, which may not resonate as deeply with cultural narratives.

Destructive bots across both movements, however, tended to use cultural resonance much less frequently than constructive bots. This preference is likely because as mentioned in the above section, emotional appeals can disrupt and polarise discussions more immediately, while cultural resonance does not have such immediate effect, so destructive bots focus more on the former than the latter.

Social bots use cultural resonance to shape the DOS of activists both directly and indirectly. As mentioned, cultural resonance ensures that specific messages and actions align with and evoke responses based on shared cultural values, symbols, and identities. Directly, they generate emotional reactions from a third-party audience, thereby impacting the DOS of activists. Indirectly, social bots aim to achieve cultural resonance by constructing common and trustworthy identities to make their messages more credible and persuasive.

Both pro-protest (constructive) and con-protest (destructive) bots first directly appeal through common cultural values and symbols. Constructive cultural resonance explicitly calls upon or implicitly conveys that protesters are resonating with cultural values like equality, racial justice, kindness, and solidarity. Conversely, destructive cultural resonance frames anti-BLM and anti-XR protestors as 'freedom fighters,' emphasises the right to protest as a '1A right,' (BLM2, #13)

and portrays those who disapprove of the radical activities of XR as sane and possessing common sense.

Furthermore, both constructive and destructive bots attempt to construct identities that resonates with common human experiences, so as to indirectly achieve cultural resonance, but the identities they chose often differ. Nevertheless, for both parties, identities with cultural resonance increases the credibility and persuasiveness of their messages and consequently encourages third-party engagement.

Constructive bots achieve resonance by openly identifying themselves as automated agents. Consequently, they aim to achieve cultural resonance through their language use. For example, a bot account designed to help people petition against police violence automated the service of writing to senators. In promoting this service to a general audience, it used phrasing like 'write to your officers' to create a sense of civic duty and responsibility, thereby resonating with cultural values of community engagement and activism. Furthermore, operators of the bots suggest that people can sign the 'Move to End Police Brutality' petition, indicating that the bot account supports this cause. This approach makes it easier for them to gain support from potential pro-protest audiences who can be mobilised for protest.

Destructive social bots, on the contrary, fabricate online identities that appeal to human users to achieve cultural resonance. These bots construct fake personas through their profile details, creating a sense of 'sameness' with humans. For example, these personas often include specific personal elements like family members, pets, and hobbies to appear authentic. One automated account described itself as a 'father of two kids and dog lover,' and another similarly portrayed itself with personal details such as hobbies or community involvement. At least a hundred users in our BLM bot database described themselves as 'dog lovers.' The purpose of these fabricated personas is to gain trust and foster a connection with real users, making their disruptive messages more likely to be accepted.

Furthermore, destructive bots also adopt context-specific identities relevant to protests to increase their resonance. For instance, during the BLM protests,

some bots pretended to be African Americans criticising protest goals, incorporating stereotypical elements related to gender and race to appear genuine. This strategy enhances their credibility and the likelihood that their messages will be accepted and shared, affecting the discourse opportunity structure for activists. Human users, particularly activists, often identify these deceptive accounts and raise awareness. For example, a suspected bot spotted by activists (USR1 #156) had 'BLM' in its account name and tweeted about relationship insecurities, using gender stereotypes to appear as a black woman. This account then tweeted counter-slogans like '#AllLivesMatter,' which are seen as dismissive of the issues addressed by '#BlackLivesMatter.' An activist reported (RD1 #11) seeing a similar account on Instagram using a black man' s profile picture to comment, 'All Lives Matter,' on BLM posts.

In summary, both forms of resonance strategies were observed in bot activities, with constructive bots focusing on cultural resonance, especially in the BLM context, while destructive bots preferring emotional resonance to quickly disrupt discussions. Either way, these bots used various narrative strategies to engage with users and influence discourse, ranging from provocative messaging to adopting context-specific identities that resonate with their target audience. While past research has primarily focused on emotional resonance, our findings reveal that cultural resonance is equally common among both con- and pro-protest bots, and gives rise to new strategies that have not been previously identified.

### 6.4.3 Legitimacy

Legitimacy in our DOS framework refers to the level of support for activists' claims, expressed through endorsement of protest frames, support for activists, and participation in online activism within the discourse. The interaction between bots and humans collectively shapes the legitimacy of protests in these three dimensions through various distinct strategies. As with the resonance section, we will first provide an overview of the percentage of bot activities targeting those specific

**Table 6.5:** Distribution of Legitimacy Categories in Bot Activities for BLM and XR Datasets

| Bot Type | Legitimacy Category | % of Same Bot Type (BLM) | % Overall (BLM) | % of Same Bot Type (XR) | % Overall (XR) |
|---|---|---|---|---|---|
| **Constructive** | Legitimacy of Protest Frame | 69.13% | 46.8% | 51.42% | 21.7% |
| | Image of Activists | 25.01% | 16.93% | 22.75% | 9.6% |
| | Protest Participation | 33.23% | 22.5% | 52.37% | 22.1% |
| | Not Applicable | 0.44% | 0.3% | 5.45% | 2.3% |
| **Destructive** | Legitimacy of Protest Frame | 51.39% | 16.6% | 30.62% | 17.7% |
| | Image of Activists | 75.54% | 24.4% | 47.58% | 27.5% |
| | Protest Participation | 11.76% | 3.8% | 15.92% | 9.2% |
| | Not Applicable | 0.62% | 0.2% | 9.34% | 5.4% |

Note: The percentages represent the distribution of messages for each bot type across specified categories in the BLM and XR samples. "Same Bot Type" refers to messages from the corresponding bot category (e.g., 69.13% of all messages from constructive BLM bots aim to increase the legitimacy of protest frames). "Overall" indicates the distribution of messages across all bot categories within each protest. Since legitimacy categories can overlap, a single bot message may use multiple legitimacy strategies, meaning the overall percentages may exceed 100%.

dimensions within a selected sample of astroturfing bot records. We then detail bot activities aiming to impact those sub-dimensions with specific examples.

All tactics are widely adopted by social bots in both protests, whether constructive or destructive, with only about 5% of astroturfing records not targeting the legitimacy of the protest. A detailed breakdown of the percentage of tactics used by constructive and destructive bots across the two cases is presented in Table 6.5. This high prevalence shows the strategic importance of legitimacy in shaping public perception of protests. Both constructive and destructive bots focus on influencing legitimacy to either bolster support or undermine the protest' s credibility, which reflects their intent to maximise impact on public discourse. By targeting legitimacy, bots can effectively affect how protests are framed and perceived, and even impact the mobilisation and outcome of protests.

### Legitimacy of protest frames

Both constructive and destructive social bots actively engage with protest frames to influence online activism. Constructive bots focus on enhancing the legitimacy

of the protest. For BLM, 69.13% of messages from constructive bots (within pro-protest records) and 46.8% of all BLM-related messages emphasised legitimising the protest framework. Similarly, for XR, 51.42% of constructive bot messages (within pro-protest records) and 21.7% of all XR-related messages focus on this aspect. These high percentages show that constructive bots are deeply committed to reinforcing the legitimacy and support for the protest by highlighting and validating the protest' s goals and messages.

To impact legitimacy, pro-protest bots in the two cases employ different tactics tailored to the specific goals of each movement. In the BLM context, bots emphasise why individuals should participate to voice their grievances about Floyd's death, often highlighting the moral imperative and immediate need for justice. In contrast, XR bots emphasise the urgency of addressing climate change with posts. This difference is partly because racial justice issues, like those highlighted by BLM, are already well-recognised and require reinforcement of existing support. BLM pro-protest bots call upon immediate action against the case of George Floyd, and therefore focus on the legitimacy of this specific case in their protest frame. The climate change narrative promoted by XR is still evolving and faces ongoing skepticism, necessitating a focus on justifying its urgency.

Destructive bots also engage in impacting the legitimacy of protest frames, and they exhibit a varied approach tailored to each protest movement' s context. For XR, 30.6% of messages from destructive bots (representing 17.7% across all XR records) targeted the legitimacy of the protest framework. Their narratives aimed at delegitimising protest frames cast doubt on the relevance and impact of climate change activism itself. Given the multifaceted nature of climate activism, these bots highlighted perceived contradictions within the global climate change landscape or questioning the effectiveness of its tactics. Examples include statements like, 'Go protest climate change in China and India, the two countries severely polluting the environment' (XR2, #221) and 'Like everyone in that stadium has never heard of climate change, right? What a useless protest' (XR2, #214).

In contrast, destructive BLM bots demonstrated a greater focus on questioning the legitimacy of protest frames than their XR counterparts. Approximately 51.39% of their messages were dedicated to this tactic, comprising 16.6% across all BLM records. These bots specifically targeted the goals and credibility of the George Floyd protest, likely due to the case's high-profile nature, which made it a focal point for public discourse. Given the widespread support for the racial justice movement, bots may have sought to exploit the heightened attention and polarising views around the specific case to undermine its legitimacy. These bots propagate narratives that questioned the authenticity of the BLM movement, often using misleading content. For instance, one message promoted a documentary claiming, 'A documentary revealing the truth behind BLM! Probably what all of us Freedom Fighters already know!!' (BLM1, #426), which purported to expose fabrications surrounding George Floyd' s death, thereby aiming to discredit the movement's motives and erode public support.

**Image of activists**

Legitimacy also involves shaping the public image of activists, and social bots play a significant role in influencing this perception, especially destructive ones. By discussing activists' behaviours and moral standards, these bots aim to either bolster or undermine public support for the protests.

A smaller proportion of constructive bots focus on positively portraying activists, compared to destructive ones. The number is similar across the constructive bots in the two cases: this accounts for 25.01% of pro-protest bots in the BLM sample and 16.9% across all BLM records in our sample, as well as 22.75% of pro-protest bots in the XR sample and 9.6% across all XR records in our sample. This limited focus is likely because constructive bots primarily prioritise amplifying protest messages and building support through other strategies, such as enhancing the legitimacy of the protest frame or resonating emotionally with their audience.

Constructive bot activities that focus on the image of activists work to establish a narrative that the online group of activists are cohesive and full of solidarity.

As such, we observed several cases where social bots actively worked to increase solidarity and cohesiveness among activists. These bots first share existing records that demonstrate cohesiveness and solidarity within protest activities and communities. For example, one bot frequently posts links like 'Moments of kindness and solidarity shine during George Floyd protests.' (BLM2, #353) Additionally, bots use automation to trace pro-protest posts on the internet and offer encouragement. A bot developer mentioned creating a bot that 'leaves a comment on any solid images found in the #blacklivesmatter tag.' (RD1, #7) When a protest movement shows high levels of cohesion and solidarity, it appears more legitimate and credible to outsiders. Therefore, this type of engagement fosters a sense of community and support among activists, reinforcing the perception of a unified movement.

In contrast, a much larger proportion of destructive bots focused on attacking the image of activists. About 75.5% of destructive BLM bots focused on the image of activists, which accounts for 24.4% of all BLM bot records in our sample. For XR, 47.58% of destructive bots targeted negative aspects of activists' images, making up 27.5% of all XR bot records in our sample.

Destructive bots in both cases diminish the image of activists by questioning the motivations behind protests or portraying activists' behaviours as extreme. These bot messages suggests that activists' personal actions contradict their public message, and effectively disrupts and challenges the growing public acceptance of BLM and XR. In the case of BLM protests, these messages achieve this by claiming that BLM protests have gone beyond acceptable boundaries, such as rioting or looting during Floyd protests. For XR protests, disruptive bots initially spread news or conspiracies aimed at highlighting perceived hypocrisy among protestors. For example, they may astroturf claims suggesting XR is supported by fossil fuel companies. Additionally, they criticise actions like activists using private jets, which appear contrary to their environmental message: 'You Protest. Then hop on your Private Jet to your Riches!' (XR1, #188). Similar to disruptive bots in the BLM case, they also argue that XR activists disrupt public order, citing instances where activists glued themselves to buses, causing inconvenience.

**Encouraging or weakening participation**

Protest participation is another essential aspect of DOS legitimacy, making it also a key focus for both pro- and con-protest bots.

Pro-protest bots in both cases focused on encouraging people to join the protests, with a higher percentage observed in the XR case. 33.2% of pro-protest bots associated with Black Lives Matter (BLM) focused on boosting participation. This represents 22.5% of all BLM-related bot activity in our astrouturfing sample. In comparison, 52.37% of pro-protest bots related to Extinction Rebellion (XR) are aimed at increasing protest engagement, which makes up 22.1% of all XR-related bot activity in our sample. Mobilising supporters is crucial for making movements more visible and impactful. Large, disruptive actions, which are often necessary for these movements, depend on high levels of participant turnout and online engagement to be effective.

To achieve this, pro-protest bots act as knowledge-sharing tools, providing instructions on how to participate. For example, one pro-protest bot might share a message like, 'From donating money to attending a protest, here are some things you can do to stand up to racial injustice' (BLM2, #38), along with a link to an article. These bots use hashtags like #BLM to spread detailed instructions on participating in BLM, including posting in celebrities' comments. Additionally, some bots lower the cost of action through automation, sending messages like (BLM1, #1323): 'Text POLICE to [NUM] to find police violence data in your town and state, and then sign...to have it delivered to your state officials.'

Activities aimed at impacting protest participation were less common among con-protest bots compared to pro-protest bots, but they still accounted for a small proportion of their overall activity. Specifically, 11.76% of con-protest bots related to Black Lives Matter (BLM) focused on weakening participation, which accounts for 3.8% of all BLM-related bot activity in our sample. For Extinction Rebellion (XR), 15.92% of con-protest bots targeted reducing participation, representing 9.2% of all XR-related bot activity in our sample.

The smaller overall proportion for destructive bots was possibly because their tactics, such as spreading fake news, were less visible. Instead of overtly discouraging participation, these bots often operated in the background, subtly undermining protest efforts by sowing doubt, fear, and confusion. While constructive bots aim to inform and encourage participation, con-protest bots undermine these efforts by spreading false information. During the George Floyd protests, there were reports of protesters encountering bot or troll accounts posting about fictitious offline events. For instance, bots falsely claimed there would be a racial justice gathering at a non-existent location, which was quickly debunked as fake (USR1, #44). This indirect approach made their activities less obvious but still effective in weakening participation and the overall impact of the movements.

There is limited speculation among protest organisers about the motives behind these actions, but according to our framework, a plausible interpretation is that bots aim to confuse the diffusion of protest-related information online. By promoting fake protest events, bots not only hinder the recruitment for genuine protests but also erode trust in protest organisations among the general audience. These activities collectively contribute to diminishing the legitimacy of protest frames.

It is also worth mentioning that strategies that impact legitimacy dimensions - the legitimacy of protest frames, the image of activists and the extent of participation - are often combined with those in the 'visibility' section to enhance exposure. For example, a pro-protest bot may reach out to people with a #BLM hashtag in their post (targeted engagement) to express solidarity to enhance people' s view of online activists, and a con-protest bot may orchestrate (astroturfing) to spread protest-related conspiracy to negate people' s view on protest frames.

Finally, based on our newly revised DOS framework, we discovered that bots play a important role in shaping protest legitimacy by influencing public perception, manipulating activist images, and directly impacting protest participation. Additionally, we uncovered several new strategies, such as the use of fake news to undermine participation, that have not been widely discussed before. These

overlooked tactics are essential to understanding how bots can significantly affect the success and dynamics of social movements.

### 6.4.4 Overview of social bots' role in the two protest cases

In the last three sections, we illustrated the behaviour and strategies of social bots, focusing on how they either constructively support or destructively undermine the discourse of activists. Now, we provide an overview of the roles social bots play in two cases: XR and BLM.

Overall, the behaviour of social bots has significantly influenced the visibility, resonance, and legitimacy of activists' discourse, shaping how protest movements are perceived and engaged with. In terms of visibility, for instance, we previously noted that 25% to 30% of bot messages were dedicated to astroturfing. These bots enhance the visibility of both constructive and destructive messages in relation to protests in both cases. Regarding legitimacy and resonance, 49.6% of unique bot records for Extinction Rebellion (XR) and 88.3% for Black Lives Matter (BLM) specifically targeted resonance, while 95% of these records sought to impact legitimacy.

Although this research does not quantify the exact influence of bots, it is evident that protests involving bots differ significantly from those that occurred in earlier stages of social media, before bots became widespread. Social bots have introduced new dynamics by amplifying certain messages, shaping the visibility of pro- and anti-protest narratives, and influencing public perceptions. This has altered how protests are organised, how information spreads, and how movements gain legitimacy. Even without precise quantification, it is clear that the presence of bots has significantly transformed the protest landscape compared to earlier, pre-bot protests.

While both constructive and destructive bots employ various strategies across the two cases, the specific focus of these strategies appears to be tailored to each movement. This distinction is crucial because the dynamics and effects of bot activity differ significantly between XR and BLM, likely due to the unique goals,

**Table 6.6:** Overview of bot activities in the DOS of the two cases

| | **Constructive bots** | **Destructive bots** |
|---|---|---|
| **XR** | - Less common<br>- Elicit positive emotions such as optimism and hope<br>- Reinforce the legitimacy of climate change activism<br>- Encourage protest participation | - Make up the majority (67%) of XR bot activity<br>- Primarily spread anti-protest messages<br>- Undermine XR by attacking activists' credibility (e.g., hypocrisy)<br>- Less focus on emotional appeal (resonance) |
| **BLM** | - Make up the majority (71.9%) of bot activity<br>- Linked to fan-based activism<br>- Enhance cultural resonance about racial justice<br>- Strengthen the legitimacy of BLM protest narratives in all aspects | - Less common than constructive bots<br>- Push other political agendas, such as support for Trump<br>- Trigger strong emotions, especially anger<br>- Discredit BLM legitmacy across various dimensions<br>- Particular emphasis on tarnishing the image of activists |

contexts, and nature of each movement. Table 6.6 highlights how constructive and destructive bots function differently in relation to XR and BLM. We will summarise these differences and explore why bots act in contrasting ways in these two contexts.

In the case of XR, destructive bots dominate, making up 67% of the recorded activity. These bots primarily focus on spreading anti-protest messages, increasing the visibility of dissenting voices while diminishing the reach of activists' messages. They concentrate more on challenging the legitimacy of XR compared to resonance, particularly by attacking the credibility of activists, often through claims of hypocrisy. This focus on legitimacy may be tied to how climate change activism, which XR promotes, is frequently framed as controversial or disruptive. The goal of these bots could be to discredit activists as extreme or inconsistent, thereby undermining the entire movement. Emotional appeal, such as anger or fear, plays a smaller role here, likely because climate change discussions are often framed more in terms of logic and evidence rather than immediate personal emotions.

Constructive bots in XR, although less prevalent, still account for about one-third of the records. In terms of resonance, their primary focus is on evoking emotional resonance, generating positive feelings like optimism and hope. These bots aim to legitimise climate activism and, crucially, encourage direct protest participation. The emphasis on motivating action—rather than simply spreading messages—may be because XR is a movement centred around physical, visible non-violent disruptions in the public sphere like road blockades. In contrast, such a focus on protest participation isn' t as evident in the BLM context, possibly because of the different forms of activism involved.

In the BLM case, the roles reverse, with constructive bots making up the majority (71.9% of the astroturfing records). One likely reason for this is the involvement of fan activism, particularly from K-pop fans, who used social media platforms and bots to promote fundraising activities. These constructive bots focus primarily on enhancing cultural resonance, linking racial justice to broader societal themes of equity, identity, and historical justice. Given that BLM' s fight against racial injustice is deeply intertwined with cultural and social issues, these bots help amplify messages that resonate within the cultural context of race, community, and shared history. The cultural resonance BLM taps into is perhaps broader and more emotionally charged than the environmental concerns raised by XR, which may explain why the bots in the BLM case are so successful at promoting the movement's narrative across different social and cultural lines.

Destructive bots targeting BLM, in contrast, focus heavily on pushing alternative political agendas, often aligning with right-wing figures like Trump. In terms of resonance, these bots are designed to provoke strong emotional reactions, particularly anger, by stoking racial tensions or attacking the legitimacy of BLM from a political angle. The emotional nature of racial justice, which directly impacts personal identity and community belonging, makes it fertile ground for these bots to manipulate sentiments and fuel division. In terms of legitimacy, by focusing on the image of activists, particularly in ways that discredit their integrity or intentions, these bots seek to undermine the movement' s credibility and reduce public support.

In summary, the differing roles of bots in Extinction Rebellion (XR) and Black Lives Matter (BLM) demonstrate the unique focus and dynamics of each movement, which ultimately shape the strategies of social bots involved. In terms of visibility, XR experiences higher levels of con-protest activity that dilutes discussions through irrelevant content, whereas BLM sees more pro-protest bot amplification but also more direct engagement from opposing bots. In terms of resonance, pro-protest bots supporting XR focus on emotional resonance to drive participation, while BLM bots rely more on cultural resonance to embed the movement within broader

societal narratives. At the same time, con-protest bots dismiss XR' s urgency and use emotional resonance against BLM to deepen polarisation. In terms of legitimacy, pro-protest bots reinforce protest framing more strongly in BLM, while XR bots balance their efforts across framing and participation. Con-protest bots primarily attack activist image, portraying BLM figures as violent and XR activists as hypocritical. XR also faces more distraction tactics, while BLM is undermined through direct counter-messaging.

These findings indicate that social bots tailor their strategies and behaviours to align with the specific characteristics and goals of each movement. Moreover, they also show that our understanding of social movements is incomplete without acknowledging the impact of social bots, as they actively shape narratives, influence public opinion, and mobilise support. Without considering the role of these digital entities, we miss an important aspect of how movements operate and interact in the digital landscape, ultimately limiting our comprehension of the strategies and challenges faced by contemporary activism.

## 6.5 Discussions and conclusions

In this article, we have examined how social bots influence the DOS of activists online. To achieve this, we refined the existing DOS theory by adjusting its dimensions and introducing relevant sub-dimensions. With this enhanced framework, we empirically investigated the impact of social bots on the DOS of activists through their automated behaviours and discursive strategies. Additionally, we analysed the scale and extent of bot activities using a sample of their interactions. In this section, we will discuss the findings, contributions and implications of this research.

We first contribute theoretically by revisiting the original DOS framework by Koopmans et al.[326] and adapting it for the contemporary context of online activism. The original framework of DOS focused primarily on offline protests and the role of traditional media. Based on past work[5, 251], this research acknowledges the hybrid nature of current civic political participation, which involves a complex media environment with diverse agents[15], including both humans and bots. We

recognised that discourse functions both to disseminate information and as a form of activism in this context. Based on these new developments in online activism, we refined the definitions of DOS dimensions—visibility, resonance, and legitimacy—to better suit the digital context. This updated framework provides a new, replicable analytical tool for assessing how various agents, including social bots, affect the emergence, mobilisation, and evolution of protests through discourse, especially on social media.

Based on this framework, we contribute empirically by assessing the extent of bot activities using a sample of bot activity records. We identify more than 9,000 unique automated-like tweets dedicated to astroturfing or targeted engagement—strategies designed to increase visibility, encompassing both constructive and destructive roles. With our sample astroturfing records, we quantify the proportion of these strategies used for visibility, resonance or legitimacy.

Empirically, we first introduced and systematically conceptualised the strategies employed by social bots, using the DOS framework to categorise these strategies across three key dimensions: visibility, resonance, and legitimacy. This approach is important because previous research on social bots has been fragmented, often focusing on identifying just one strategy at a time without offering a comprehensive understanding of how bots operate as a whole. Consequently, we lacked a full picture of what social bots do in the context of protests. However, our findings provide a thorough understanding of bot strategies and their broader impact on activist movements, using two specific cases. These bot strategies attempt to influence how protest frames are accepted and disseminated among the public and shape the overall discourse opportunity structures of activists online.

Through detailed analysis and empirical examples, we explored how bots manipulate visibility, resonance, and legitimacy to affect the spread and reception of activist messages. Overall, bots made full use of identified strategies such as astroturfing, but we have also observed new strategies that emerged in the context of XR and BLM. For visibility, bots enhance the reach of activist messages by automating large-scale broadcasts, synchronising posts for greater impact, and

targeting influential users to amplify key narratives. In terms of resonance, bots craft messages designed to evoke emotional responses and to enhance cultural resonance, while also adopting relatable or deceptive personas, such as pretending to be everyday users or protest participants, to build cultural resonance with a broader audience. Regarding legitimacy, bots manipulate the public perception of protest frames and image of activists, and even encouraging or discouraging actual protest participation.

Another key contribution of our study is the quantification of bot activity within our framework and the analysis of how these strategies are applied across the two cases. In terms of visibility, XR faces higher levels of con-protest bot activity that dilutes discussions with irrelevant content, whereas BLM experiences more pro-protest bot amplification alongside greater direct engagement from opposing bots. Regarding resonance, pro-protest bots supporting XR prioritise emotional appeals to drive participation, while those in BLM focus on cultural resonance to embed the movement within broader societal narratives. Con-protest bots, however, undermine XR by dismissing its urgency and exploit emotional resonance in BLM to intensify polarisation. For legitimacy, pro-protest bots in BLM reinforce the protest frame more strongly, whereas XR bots distribute their efforts between framing and participation. Con-protest bots primarily target activist image, portraying BLM figures as violent and XR activists as hypocritical. Additionally, XR is more susceptible to distraction tactics that shift focus away from climate activism, while BLM is more frequently subjected to direct counter-messaging designed to erode its legitimacy.

Such compare and contrast highlights the significant role that bots play in shaping the discourse surrounding activism in the digital space, particularly in the two cases of Extinction Rebellion (XR) and Black Lives Matter (BLM). This has major implications for our understanding of online activism. In the field of social media, it becomes clear that a comprehensive analysis of activism cannot be achieved without considering the influence of bots. Their strategic interventions

affect discourse on both sides—supporting and opposing activist movements—and are tailored to the specific dynamics of each case.

This finding suggests that social bots are not merely passive or disruptive actors, but active participants in shaping how activist messages are framed, received, and debated online. As mentioned, it is clear that their involvement has significantly altered the dynamics of protests compared to those in earlier stages of social media, before bots became widespread. Even though a precise quantification of such impact is beyond the scope of this research, it is evident that the presence of bots has significantly reshaped the protest landscape. As a result, future research must first take these digital agents into account to fully understand the mechanisms that drive online activism. Furthermore, future research should aim to quantify these influences, perhaps by conducting counterfactual investigations that model the impact of protest discourse without bot involvement. Such approaches could provide a more concrete understanding of how bots transform online activism.

In particular, there are two major features of social bots worth noting and further investigation. First, we discovered a substantial number of suspected to be automated agents constructive to online activism, a finding that contrasts sharply with the focus of previous research, which largely overlooked their existence. This highlights and fills an important gap in the literature, where past work has been limited to identifying the destructive roles of bots, without considering their constructive potential. This contrasts with existing research [18, 62], which often focuses on how automated accounts spread misinformation and inflammatory content that undermine activist messages. Instead, we found that a significant proportion of bot activities can actually enhance the visibility, resonance, and legitimacy of activist causes. This discovery is particularly important because it broadens our understanding of the complex role social bots play in shaping online discourse. Instead of seeing bots as purely negative or harmful, this study suggests that they can contribute positively to activist efforts, and future research should address this aspect of social bots.

Second, we also compared and contrasted the strategies used by social bots in the two cases—Extinction Rebellion (XR) and Black Lives Matter (BLM)—and found significant differences. These variations in the use of DOS strategies by bots stem from the distinct nature of the two protests. By presenting an overview of bot activities across both cases, we can see how bots adapt to the cultural, emotional, and social dynamics unique to each protest. Future research should take into account the fact that bots tailor their approaches to specific movements and contexts, recognising that their influence is shaped by the particular goals, social dynamics, and vulnerabilities of each protest. This suggests that a generic approach to understanding social bot activities is insufficient; instead, we need to explore how bot activity varies across different movements and tailor our research methods accordingly.

Beyond those theoretical and empirical contributions, this research also has significant practical implications. By uncovering and classifying the ways in which social bots influence the DOS for activists, this study provides valuable insights for both activists and policymakers. Activists can use this knowledge to enhance their strategies for engaging with online platforms and counteract malicious bot activities, while policymakers can better understand the dual role of social bots in digital discourse. This understanding is essential for all stakeholders in online activism to develop effective measures to unleash the positive potential of AI in activism while mitigating its risks.

Finally, this study has two major limitations. The first limitation of this study stems from its reliance on data from two protest cases in 2019 and 2020. This temporal focus may limit the representativeness of the findings regarding the broader landscape of protest and social bot activities. Furthermore, the social bots analysed in this research were before the introduction of advanced large language models (LLMs), and the evolving nature of these technologies could influence bot behaviour in new ways[93, 361]. Despite these constraints, the findings remain highly relevant. The theoretical framework developed in this study is designed to be replicable, providing a solid foundation for future research. Furthermore,

the fundamental characteristics of automation identified in this study continue to be applicable. Nevertheless, this research is exploratory, and it highlights and conceptualises critical discursive strategies employed by bots. Such exploration emphasises the need for continued investigation as bot sophistication increases with the advancement of LLMs. As such, the insights gained here offer valuable starting points for understanding and addressing the evolving challenges posed by social bots.

The second limitation is related to the scope of the study, which focused exclusively on bot activities. While the analysis considered the context of human interactions and platform mechanisms, these aspects were not the primary focus of this research. Future studies should address this gap by exploring how human factors influence responses to bots and investigating the impact of platform mechanisms on bot design. Such research will provide a more comprehensive understanding of the dynamic between bots, human users, and digital platforms, offering deeper insights into the broader dynamics of online discourse and activism.

# 7 Discussions and conclusions

## Contents


**7.1 Findings . . . . . 184**
7.1.1 Background of research . . . . . 185
7.1.2 Bot detection methods pipeline . . . . . 186
7.1.3 Bots shape micro-level human behaviour . . . . . 187
7.1.4 Bots contribute to a decline in cohesiveness within human networks . . . . . 188
7.1.5 How bots target visibility, resonance and legitimacy of protests . . . . . 189
**7.2 Contributions of the dissertation . . . . . 190**
7.2.1 Theoretical contributions . . . . . 190
7.2.2 Empirical contributions . . . . . 192
7.2.3 Methodological contributions . . . . . 194
**7.3 Practical implications . . . . . 195**
**7.4 Limitations . . . . . 198**
**7.5 Future research directions . . . . . 201**
**7.6 Conclusions . . . . . 205**


This thesis examines the behaviour and impact of social bots in online activism and related political communication, focusing on two case studies: Extinction Rebellion (XR) and Black Lives Matter (BLM). Online activism is profoundly shaped by social media, which influences the diffusion of information, the formation of social networks, and the mobilisation of participants. Social bots are a significant factor affecting the social media public and micro-level human behaviour

in online activism.

The four interconnected empirical chapters addressing this research topic begin by establishing a systematic methodology to detect social bots and explore their potential impact. The chapters then offer empirical insights into the potential causal relationships regarding how social bots affect human behaviour in communication and human-to-human connections online. Using qualitative methods, the final empirical chapter explores the mechanisms by which bot strategies and user perceptions of social bots induce changes in human behaviour. Overall, these results offer novel insights into how social bots influence human behaviour in the context of online activism through systematic investigation. Consequently, this research sheds light on topics such as social movement studies, micro-level behaviour in political communication, the societal impact of AI, and other broad sociological issues. It also has profound practical implications for activists, policymakers, digital media platforms, and the general public.

In the following sections of this chapter, I will first recap the conclusions of each empirical chapter and then discuss how these connect to illustrate a broader picture of the impact of social bots on human behaviour in online activism. This chapter will then outline the epistemological contributions and practical implications of this dissertation. Following that, a discussion of its limitations will be provided. Finally, this chapter will suggest further research directions based on the contributions, implications, and limitations of this dissertation.

## 7.1 Findings

Overall, this research provided a coherent and complete framework to address how social bots impact human users in the context of online activism on social media. This section provides a overview of the key findings.

This dissertation began by outlining the background of research and identifying key gaps in the study of social bots and their role in online social movements. It then provided a comprehensive literature review relevant to the main topic, followed

by the development of a robust methodology for detecting automated agents on social media and assessing their impact on human behaviour.

Building on this foundation, the study found that social bots influence human behaviour at both the micro and meso levels. At the micro level, users who interacted with bots in the context of protests exhibited changes in sentiment and posting frequency compared to those who did not. At the meso level, interactions with social bots were associated with a decline in social cohesion within online activist communities. The dissertation then explored the mechanisms driving these changes, demonstrating that bots employ a range of strategies across visibility, resonance, and legitimacy to shape online discourse. Social bots play both constructive and destructive roles, either amplifying or undermining activist messages depending on the movement' s unique characteristics and context.

### 7.1.1 Background of research

This dissertation began (Chapter 1) by emphasising the research's importance, outlining the background, identifying gaps in current knowledge, and discussing potential contributions. It presented an overarching research question along with three sub-research questions and then outlined the structure of the dissertation.

Chapter 2 provided a comprehensive review of the literature relevant to the main topic, focusing on four key dimensions: bot studies, social media and protest, the role of social media platforms, and research on the two protest cases. The review of bot studies covered definitions, taxonomy, identification methods, bot strategies, and the potential origins of social bots. The chapter also reviewed how social media impacts online activism through three perspectives: information diffusion, online social networks, and protest mobilisation tactics. It reviewed how social media influences protests and how bots mediate these effects. Although not the primary focus of this dissertation, this chapter also reviewed how social media platforms affect the activities of both activists and social bots, as social media platforms provide essential context for both activities. Special emphasis was placed on Twitter (X), the primary platform where the case studies occur and

from which we collect our data, to justify its selection as the focal point of our research. Furthermore, the chapter explored past research on the two protest cases —Extinction Rebellion (XR) and Black Lives Matter (BLM), so as to put the above elements in the context that they impact the dynamics of online activism. In each dimension, the review highlighted research gaps and set the stage for the study's contributions.

### 7.1.2 Bot detection methods pipeline

Chapter 3 established a robust methodology for identifying automated agents on social media and examining their impact on human behaviour. It serves as the foundation for addressing the central research question (RQ)-How do automated accounts affect online collective actions and related discussions on social media-and the three main RQs. The research combined existing open-source tools with self-developed supervised machine learning methods to identify social bots. It included robustness checks for classification thresholds (e.g., classifying users with a bot probability above the threshold as 'bot,' and the rest as 'human' or 'unknown') to minimise false positives and false negatives. The chapter also covered data collection, natural language processing, text classification with large language models, and approaches to potential causal inference, which served as a methodological foundation for the empirical chapters.

This research situated itself within online political communication, and therefore required a thorough understanding of protest-related discourse, opinions, and trends. This chapter then developed a common data collection, natural language processing, and inference pipeline to identify patterns and themes in both XR and BLM cases. This pipeline supported the subsequent empirical chapters. To quantify the influence of social bots on human behaviour, the research formulated a structured research design pipeline used in Chapters 4 and 5.

### 7.1.3 Bots shape micro-level human behaviour

With these methodological setups, Chapter 4 investigated how bots shape micro-level human behaviour using the case of XR, hence addressing RQ1: How social bots influence human communication behaviour in protest-related discourse?

This research first found that social bots and human users together generated information cascades in discussions around climate change protests. These cascades often involved heated discussions of specific protest events or upsurges of anti-protest sentiment. In all identified cascades, the number of bot tweets during the information cascade could predict human activity, suggesting that social bots were diverting human activities in online political communication.

Moreover, with a natural experiment approach, the chapter revealed that interactions with social bots had a detrimental effect on human users' sentiment and posting behaviour. This chapter compared communication histories between human users who directly interacted with bots and matched human users who did not. The findings demonstrated a consistent negative impact of bot interactions on subsequent human sentiment, with exposed users displaying significantly more negative sentiment than their counterparts. The nature of bot interaction also influences human tweeting activity and sentiment toward protests. Political astroturfing bots increase activity, whereas other bots decrease it. Sentiment changes toward protests depend on the user's original support level, indicating targeted manipulation.

These findings suggest that social bots can significantly alter public discourse by amplifying or dampening specific sentiments and behaviours. It is worth noticing that the potential role of other unobserved factors in the process remains open questions. Nevertheless, this highlights the need for greater awareness and regulation of bot activities on social media platforms. Additionally, the targeted nature of sentiment manipulation implied that bots can exacerbate polarisation and influence public opinion on contentious issues like climate change and social justice movements.

### 7.1.4 Bots contribute to a decline in cohesiveness within human networks

Chapter 5 shifted the focus to the meso-level. It investigated the impact of bots on human connections in online social networks during protests, using the case of BLM, to address RQ2: How do social bots impact the network structure of online social networks in protests?

This chapter first found distinct network structures formed by social bots that clustered around varying motivations in BLM. Previous research indicated that bots form distinct network structures, like dense 'botnets.' This study found similar structures formed by social bots in online networks, clustered by different protest support categories. During Black Lives Matter (BLM) discussions, bots coordinated to spread anti-BLM and highly partisan messages, as well as conspiracy theories and spam to disrupt protest mobilisation. Pro-protest activists and neutral entities, like K-pop fans and news media, also used bots to share protest news and fundraising campaigns. Such analysis revealed that automated accounts are active across various BLM communities, spreading diverse messages.

In this polarised, fragmented network, social bots, especially anti-protest astroturfers, caused a decline in human communications within online social networks. This effect is observed at both the individual (ego network) level and the community level. Furthermore, bots' effects on human-human networks vary significantly based on pre-existing opinions and the stance of social bots. Pro-protest users, compared to neutral or anti-protest users, seem to be most strongly affected by this effect.

These findings highlighted how social bots can fracture and polarise online communities, reducing meaningful human interaction and potentially undermining social cohesion. This emphasises the need for better detection and mitigation strategies for bot activities on social media platforms. Additionally, the varying impact of bots based on user support group suggests that interventions should be tailored to different community dynamics. Understanding the role of bots in spreading diverse messages across polarised networks can inform the development

of more resilient online activist networks and more effective digital communication strategies for social movements.

### 7.1.5 How bots target visibility, resonance and legitimacy of protests

Finally, building on the claims about how social bots may impact human behaviour, Chapter 6 qualitatively investigated the mechanisms behind why human users are affected by social bots, hence addressing RQ3: What are the strategies social bots use to affect discourse in online activism? Social bots exert their influence on human activists and third-party audiences through online discourse. The chapter used a computer-assisted content analysis (CACA) approach to understand bot strategies that impact the discursive opportunity structure (DOS) of activists in the two protest cases. DOS provides a framework to analyse factors that influence how activists' messages are diffused, achieve resonance, and gain support online.

Chapter 6 explored how bot strategies influence the discursive opportunities structure (DOS) of online activists. It began by refining the traditional DOS framework, originally designed for offline protests, to fit the context of online activism where social bots play a role. In this digital space, discourse serves both to spread information and as a form of activism itself, a new development that was not covered in the original framework of DOS. Through in-depth qualitative studies, the chapter then examined how social bots shape the DOS through three key dimensions: visibility, resonance, and legitimacy.

This chapter provides a comprehensive analysis of social bots' roles in the two protest cases. It shows bots use a variety of strategies across visibility, resonance, and legitimacy. It also highlights two major features of bots in the protest cases: first, social bots play both constructive and destructive roles, enhancing or undermining activist messages depending on the context. Contrary to prior research, this study identifies a significant presence of pro-protest bots that amplify visibility, cultural resonance, and legitimacy. Second, by comparing these movements, the study reveals that bot strategies differ based on each movement' s unique

characteristics. For example, XR bots focus on emotional resonance, while BLM bots focus on cultural resonance.

These findings show it is impossible to fully understand modern day activism without taking automate agents into account. This understanding is vital for future research of social movements, and for activists aiming to use AI to enhance their online influence and combat harmful bot activity. The revised DOS framework also provides researchers with a useful tool for studying interactions between automated agents and human discourse in digital activism.

## 7.2 Contributions of the dissertation

Overall, this research provides a comprehensive answer to the research question of how social bots impact human users in the context of online activism on social media. These research findings have profound contributions theoretically, empirically and methodologically to several broader areas of studies.

### 7.2.1 Theoretical contributions

This study first offers significant theoretical contributions. It is among the first to develop a conceptual framework that clarifies how automated agents influence human behaviour through online interactions on social media, particularly in the context of online activism. By expanding and refining the existing theories of social movement studies, such as the discursive opportunity structure (DOS) framework, this research adapts these theories to the modern, hybrid environment of online social media, which encompasses both human and automated agents. The traditional DOS framework, originally developed for traditional media contexts, is inadequate for addressing the complexities of today's digital activism landscape. This study effectively updates the theory to better fit the contemporary context of online activism, providing a valuable advancement in the field of social movement studies.

Such a conceptual framework also addresses the gaps in the existing descriptions of bot activity and bot typology by proposing a more precise taxonomy of bot activities in protests. This new analytical framework improves the understanding

of automated agents' roles in social movement studies and political communication. Moreover, it offers a applicable and expandable model for future research in online activism and social bot studies. Its potential for expansion makes it a valuable tool for further studies exploring the intersection of technology, activism, and social media dynamics.

Furthermore, the findings on how bots influence people's behaviour also have significant theoretical implications. Social bots, as new agents in digital environments, impact micro-level interactions and social networks online, particularly influencing processes such as information diffusion and protest mobilisation. This suggests that existing theories on collective action and hybrid media environments need to be revised to incorporate the role of bots. As mentioned in Chapter 1, the emergence of collective action presents a sociological paradox. This dissertation shows that we cannot fully examine collective actions without considering the influence of bots. Therefore, understanding the actions and impacts of bots is essential for developing theories of contemporary collective actions.

In terms of micro-level interactions and information diffusion, Chapter 2 reviewed past work concerning how social media significantly alters information diffusion patterns by lowering the costs of sharing information and changing how information spreads. Based on those past work, in Chapter 4, this study further found that social bots contribute to this phenomenon by generating information cascades—a process where information spreads rapidly through a network, often initiated by the actions of a few influential users. These bots can amplify messages that encourage mobilisation, but they can also disseminate misinformation or dissenting viewpoints that undermine activist efforts.

This dual capacity of social bots can lead to heightened public awareness and energise protest movements, but it can also contribute to polarised discussions and hinder protest mobilisation by shaping human sentiment and posting behaviour. Therefore, theories concerning information diffusion, polarisation, and collective action must be further developed to account for the influence of social bots. This

necessitates a reevaluation of how information flows in digital spaces and the ways in which these automated agents impact the dynamics of social movements.

Moreover, the role of bots in shaping online social networks (OSNs) in protests has significant theoretical implications. Chapter 2 reviewed previous research indicating that protests heavily rely on OSNs, which form distinct structures that influence protest mobilisation and foster cohesion among participants. However, findings in Chapter 5 reveal that bots may significantly disrupt these network structures, impeding genuine human communication and collaboration. These disruptions suggest that traditional theories of OSNs, particularly those related to how OSN structures facilitate protest mobilisation, may need to be revised to account for the influence of automated agents like bots.

### 7.2.2 Empirical contributions

Empirically, this research significantly advances our understanding of automated entities and their roles in online political communication and activism, which has important implications.

In the field of social bot studies, this dissertation is among the first to make claims about how automated agents affect human behaviour on social media. As social bots and AI technologies become increasingly prevalent, they significantly influence digital interactions, attracting considerable scholarly attention[247, 67]. This research systematically examines their impacts across various dimensions, ranging from micro-level human interactions to broader patterns in digital discourse, thereby providing a comprehensive understanding of how social bots shape online engagement and communication dynamics.

With both quantitative and qualitative research methods, the study investigates how social bots influence political communication and activism. This includes analysing how these bots shape visibility, resonance, and legitimacy within online discourse and explaining the underlying mechanisms of these influences. This comprehensive, mixed-method approach reveals that social bots are a significant factor in digital interaction dynamics and online activism. Additionally, unlike previous

studies that have predominantly focused on the negative aspects of automated agents[267, 248, 15], this dissertation offers a balanced perspective, illustrating how bots can both support and undermine activist efforts. This nuanced view is crucial for developing a comprehensive understanding of digital activism and political communication. By connecting the findings on the effects of automated agents to broader social phenomena, this research illustrates the multifaceted impact of social bots on social media.

Additionally, this research contributes to social movement studies by providing an empirical investigation into the roles of social bots within two significant protest movements: Extinction Rebellion (XR) and Black Lives Matter (BLM). These cases are particularly representative because they address critical issues—environmental and racial justice—within liberal democratic societies and are heavily discussed on social media. By comparing these two movements, the study offers a comprehensive view of how social bots operate across different protest contexts and political topics of interest.

The findings reveal that social bots are active in both protest cases, employing similar strategies across the DOS dimensions—visibility, resonance, and legitimacy —but with subtle differences tailored to each movement's context. Bots play both supportive and adversarial roles, influencing information flows and shaping user sentiment. In the case of XR, destructive bots dominate, negatively impacting user sentiment by undermining the movement's legitimacy and amplifying anti-protest messages, ultimately increasing discourse around the protest. For BLM, bots reduce human-to-human engagement and exhibit more politically partisan behaviour, often linked to broader events such as presidential elections. This study highlights an important distinction in bot tactics: while the overarching strategies —such as enhancing visibility and manipulating resonance—are consistent, the specific focus of each movement alters the bots' approach. For example, anti-protest bots targeting BLM are more politically motivated and push alternative agendas, while those targeting XR concentrate on discrediting the movement' s credibility.

This demonstrates that while bots employ similar strategies, their application is tailored to the unique dynamics of each protest.

These findings has three important implications. First, these empirical findings demonstrate the need for scholars in communication studies, political sociology, and the sociology of the Internet to recognise social bots as essential elements in their research on social media. Rather than neglecting or excluding these automated agents from datasets during empirical analysis, researchers should incorporate them to achieve a fuller understanding of digital dynamics.

Second, these findings not only enhance our understanding of the protest cases, but also shed light on urgent societal issues—racial justice and climate change—that have wide-reaching political, cultural, and environmental implications. By examining how social bots interact within these contexts, we gain valuable insights into how public discourse surrounding these critical political issues is shaped, distorted, or amplified in digital spaces.

Finally, these findings illustrate that comprehending large-scale social movements is impossible without considering the role of bots, as they significantly shape the dynamics of modern movements. Unlike pre-social media and pre-automation movements, today' s movements are profoundly influenced by automated agents, whose strategies, visibility, and public discourse are often distorted or amplified by bots. This underscores the necessity for future studies of social movements to account for the pivotal role that bots play.

### 7.2.3 Methodological contributions

Methodologically, this research makes a significant methodological contribution by developing a bot identification method specifically suitable to the research context. Although various bot detection tools, such as the open-source Botometer[35], exist, they have been found to be less effective in the domain of social science research, particularly in political communication and political sociology[36, 37]. This research addresses these limitations by systematically reviewing and comparing different bot identification methods in the context of social movement studies. It thoroughly

examines their explainability and auditability, leading to the proposal of a practical and replicable procedure for detecting political bots.

It is worth noticing advancements in large language models (LLMs) have introduced new tools and techniques for automating text generation. These developments also bring challenges for bot detection[93]. However, the method developed in this research remains highly relevant. LLMs can generate increasingly sophisticated and human-like text, potentially complicating the identification of automated agents[361]. Despite this, this research provides a robust framework for identifying bots by focusing on contextual and behavioural characteristics that LLMs might not fully replicate or obscure. Moreover, the method's emphasis on replicability ensures that it can adapt to evolving bot behaviours and new technological contexts.

Overall, this dissertation makes significant empirical and theoretical contributions by shedding light on how social bots may either promote or hinder the mobilisation, growth, and potential outcomes of online activism.

## 7.3 Practical implications

Apart from contributions to academic research, this dissertation also has profound practical implications in many dimensions. Overall, this research shows that social bots have a significant impact on human behaviour in online activism, both hindering and promoting it (Chapters 4 and 5). This finding is highly notable and requires stakeholders from various perspectives to be aware of it. Therefore, this study will summarise these practical implications from the perspectives of activists and other political actors, technologists, policymakers and legislators as well as platform providers. Based on this discussion, this section will also briefly address the implication of those findings for democracy in general.

This dissertation has significant practical implications for activists who aim to mobilise for social changes in the social media context, and maybe other political

actors as well. Based on its methodological contributions (Chapter 3) and qualitative work (Chapter 6), this research provides them with a replicable method and detailed description of recognising automated agents online, especially those that assume the identity of humans. Even in the age of large language models (LLMs), these methods remain accountable and replicable. Furthermore, this research offers activists potential implications, examples, and methods of using bots to facilitate information diffusion, mobilisation, and organisational strengthening in online activism (Chapter 6). It also reveals that social bots potentially cause changes in activists' sentiment (Chapter 4) and a decline in cohesiveness (Chapter 5) through various strategies, which is highly concerning for online activism in liberal democracies. However, these findings allow activists to devise strategies to combat malicious bots and misinformation, such as increasing solidarity and raising bot awareness.

These findings highlights the potential for activists to use 'good bots' to drive social change. For example, activists could make use of automation to reduce the costs of protest participation or to efficiently spread protest-related information. This offers a new avenue for enhancing the reach and impact of activism in the digital space. Furthermore, this concept could extend to other political actors involved in public welfare and civic engagement, such as using bots to share accurate information on public health, voter education, or policy issues. These possibilities highlight the need to consider the broader implications of bots in enhancing informed participation and responsible information sharing in society.

This research also has similar implications for technologists: benevolent social bot designers and technologists in social media platforms who combat misinformation. This study present a detailed description of emerging constructive bots in online activism and how they impact human emotions and cohesiveness. This can serve as valuable guidance for future AI design. Similarly, this research' s work in identifying social bots and uncovering their behavioural patterns offers guidance for social media technologists to identify and combat misinformation. For example, social bots orchestrate information cascades (Chapter 4) and create fake

identities (Chapter 6) to farm engagement. Awareness of social bots makes them more identifiable, enabling technologists to react quickly to mitigate their impact.

Furthermore, this dissertation informs social media platform providers, policy-makers, and other stakeholders on the regulation of AI systems and policy regarding social movement studies. This research finds that social bots have both negative and positive impacts on the ecology of discussion online (Chapter 4 and 5), making regulation complicated. However, this research points to several promising areas for regulation. First, it is essential to have a clear definition that differentiates between bots based on their levels of automation and their intended purposes, allowing for the regulation of harmful bots. Second, raising bot awareness is essential, as people are less impacted when they are aware of bots. Third, promoting transparency in social media platforms' bot detection and mitigation efforts can help build public trust and enhance the effectiveness of combating malicious bots.

Finally, this dissertation has profound implications for democracy, given the influence of social bots on both protest behaviour and related political communication. Civic political participation and online political discourse are a foundation of contemporary democratic expression[362, 363], and our findings highlight how social bots can shape the behaviours (Chapter 4) and human connection (Chapter 5) within them.

On one hand, those findings reveal how malicious bots, such as those linked to disinformation campaigns, can fuel populist waves that erode democratic norms by manipulating public opinion and distorting discourse. This underlines the urgent need for policy interventions and the development of a healthier online ecosystem to build resistance against these threats. On the other hand, we also find that 'good bots' have the potential to support democratic processes by aiding protests. For instance, bots can be employed to lower barriers to participation and facilitate the rapid dissemination of protest-related information (Chapter 6). This opens up new avenues for enhancing civic engagement and strengthening democratic movements. Overall, the research demonstrates the dual role of social bots in modern democracy, both as a threat and as a potential tool for democratic empowerment. These

findings make it essential to further address social bots' implications thoughtfully and strategically in the future.

## 7.4 Limitations

Despite the systematic investigation conducted in this research, it is important to acknowledge the limitations encountered during this dissertation project. Four key limitations are discussed below.

The first limitation is associated with the volatile nature of social bots. Social bots are continually evolving, influenced by changes in political participation, platform design, and technological advancements. This study focuses on social bot activity during 2019 and 2020, yet significant advancements have occurred since then. For instance, recent developments in large language models (LLMs) have made social bots more sophisticated and harder to identify, and there have been records of LLM-powered social media campaigns[93, 273]. Additionally, the study centres on bots on the platform X (formerly Twitter), which has undergone substantial changes in design and regulation[214, 213]. These ongoing changes are beyond the scope of this research and warrant further investigation.

Despite these evolving developments, the research remains relevant to current studies of AI's societal impact. It is among the first to empirically demonstrate how interactions with automated agents can affect human behaviour, including sentiment, activity levels, and social connections, and to conceptualise and empirically examine the mechanisms behind these changes. The findings and methodologies proposed are replicable and continue to be applicable to current social bots, given that the fundamental nature of automated agents remains consistent. However, future research is necessary to replicate and reassess these results in the context of LLM-driven bots and across other social media platforms.

The second major limitation concerns this dissertation' s methodological approach. This study uses a bot detection method that classifies users as 'bot,' 'human,' or 'unknown' using supervised machine learning methods. Due to the nature of supervised learning, this method cannot be 100% certain whether a user

classified as a bot is genuinely a bot[36]. Furthermore, the concept of 'bot' includes users with different extents of automation[246]. Even though this dissertation employed a tritype bot classification to address this, using fixed bot classification thresholds still struggles to capture this complexity. Many human users apply automation to increase their efficiency in posting messages or engaging with online discussions, which does not necessarily classify them as bots, yet they are no longer purely non-automated human users. Therefore, a detailed investigation of social bots and their typology is necessary, especially in the context of new technological advancements. Additionally, a fixed threshold-based approach can still lead to false positives and false negatives, potentially undermining the validity of this study' s inferences regarding bot activities. To address these concerns, this study ran those models with varied CAP scores and performed a series of robustness checks involving various thresholds for all analyses, reporting them in the main text. These steps aimed to mitigate potential misclassification and enhance the reliability of findings.

The third limitation concerns the cases and data used in this study. This research' s dataset includes several waves of Extinction Rebellion (XR) protests and the major surge in activity during the George Floyd protests, both occurring in 2019 and 2020. Although these protests are significant global events that have spanned multiple years, the focus on these specific periods may not fully capture the broader or evolving dynamics of bot activities over time. This could limit the generalisability of this study' s findings to other contexts or periods.

Nevertheless, when compared to previous studies on online activism and political communication on Twitter[25, 113],this study' s dataset provides a substantial and representative sample for examining bot activities in the context of these major protests. The use of the Twitter Academic Product Track API[213] at the time of data collection allowed for access to a comprehensive archive of tweets based on specific search queries. Given that this API is no longer available, the dataset represents one of the last extensive archives of such data, which is particularly valuable for analysis. Therefore, the dataset offers a thorough snapshot of both

bot and human activities during key protest periods, providing a robust basis for further analysis.

It is also worth mentioning that two empirical chapters (Chapters 4 and 5) focused on single cases. Chapter 4 examined the impact of bots on human behaviour using data from the Extinction Rebellion (XR) movement, while Chapter 5 explored the impact of bots on network structure using data from the Black Lives Matter (BLM) movement. While these case studies provide valuable insights, further comparative work using both datasets is needed to ensure that the observed empirical effects are consistent across different contexts. However, we have already conducted some comparative analyses; for instance, we found that information flow patterns are very similar between the two cases (See appendix A: supplementary information for Chapter 4). Additionally, Chapter 6 offers a comprehensive view of bot strategies across both cases, which helps address this limitation by providing a broader understanding of how bots operate in different protest contexts and confirming the generalizability of the findings.

Finally, the fourth limitation is that this study does not investigate the sources and origins of social bots, focusing instead on describing and quantifying their impact. Investigating the sources and origins is meaningful because it can provide deeper insights into the motivations, strategies, and resources behind social bot operations. Identifying the actors responsible for deploying social bots can help researchers understand their objectives, whether for political influence, financial gain, or other purposes. While Chapter 2 reviews existing literature on the origins of social bots, understanding who creates and operates these bots was beyond the scope of this research. Investigating the sources and origins could involve interviewing bot creators, those who combat malicious bots, or employing more complex qualitative methods. Addressing these aspects would require additional research and could provide valuable insights into the motivations and mechanisms behind social bot activities.

## 7.5 Future research directions

The central research question—'How do automated accounts affect online collective actions and related discussions on social media?'—gives rise to many possible future research directions. This question allows for potential expansion on the topic and highlights several unanswered questions that need addressing by future research. This section will summarise the potential future research directions indicated by this study.

The first direction is to extend future research to account for the current development of AI, especially the application of large language models (LLMs).This project focuses on the influence of automated accounts ('social bots') on human users, using two protest cases from 2019 to 2020. Automated accounts and the technology behind them—artificial intelligence—are rapidly evolving. LLMs are profoundly changing the design and development of social bots, thereby impacting their strategies and potential effects[93, 273]. It is therefore worthwhile to further investigate the definition and impact of automated accounts in political discussions in the context of LLMs.

Before we embark on this investigation, this dissertation has laid the groundwork to understand bots theoretically, empirically, and methodologically, serving as an essential foundation for exploring these new developments. By systematically analysing the existing literature on social bots and their interactions with users, we have identified key concepts and frameworks that inform our understanding of their behaviour and impact. Furthermore, through empirical case studies and methodological advancements, this work establishes a comprehensive baseline from which we can assess the evolving role of LLMs in shaping the landscape of automated accounts in political discourse.

Furthermore, it is essential to expand the research focus on bot impact on other political processes apart from protests. Online social media have given rise to many novel, creative, and individualised ways of political participation[290, 364]. Existing studies have made conceptual and empirical progress in understanding new forms of political participation in the era of social media. The current definition

of 'political participation' includes activities ranging from traditional voting in elections to modern expressions like signing petitions and altering one's social media profile. Social bots, or automation in general, play a role in both constructing and disrupting these methods of political participation. Past research on social bots and their impact has mostly focused on institutionalised political events such as elections[14]. By examining their impact on discussions surrounding two protest events, this dissertation serves as a solid foundation in examining the role of automated bots in protests. However, there is still a gap in understanding how automation shapes and disrupts other emerging forms of civil political participation, which future research should aim to fill.

Another promising research direction is to investigate the role of platforms in shaping social bot behaviour and compare the impact of social bots on human users across different social media platforms. This research predominantly focuses on the impact of social bots on one platform—X (formerly Twitter). We have already reviewed show our cases of research (X) shape bots' and activists' activities in Chapter 2). We also briefly mentioned how social bots make use of platform mechanisms through strategies such as targeted engagement and astroturfing (Chapter 6). However, this is not the central focus of this dissertation and therefore was not researched in detail.

Moreover, social bots operate across various platforms, not just Twitter. Furthermore, political social bots are active across a variety of social media platforms, such as Facebook, Reddit, and Telegram, where they behave and exert their impact differently[365, 66, 366]. These differences in behaviour are heavily influenced by platform design and regulations. Additionally, social movement activists rely on a wide variety of platforms to disseminate information, recruit participants, and create social change[322, 6].

This raises an important research gap: while this dissertation focuses on the impact of social bots on X in online activism, there is a need for a broader investigation into how different platforms shape bot behaviour and their interactions with users. It is still essential to further investigate the role of platforms in shaping

bot behaviour and their impact on human users. For example, exploring the role of other platforms in regulating and shaping bot design or conducting a cross-platform study to compare and contrast the effects of social bots across different platforms could yield fruitful results.

Also, while this research primarily relies on empirical data to understand how social bots impact human users in protests, future studies could benefit from experimental approaches. This research relies on field and natural experiments to raise claims about the impact of bot behaviour and uses qualitative approaches to understand the mechanisms behind those associations and potential causal claims. Experiments, especially those in lab-based settings, could provide more robust causal evidence of the impact of bots. They would also allow one to isolate specific mechanisms through which social bots influence human behaviour and examine their effect. For example, Chapter 6 finds that social bots may impact human sentiment by amplifying existing messages of humans, and with an experiment researchers could evaluate the effect of this strategy on human sentiment. This approach would therefore enhance one' s understanding of the mechanisms and factors moderating bot-human interactions in online activism.

Finally, this research sheds light on the ways in which social bots impact human users, emphasising the need for strategies to mitigate their detrimental effects or bolster their positive potential. While some implications of this dissertation have been discussed in this chapter, these insights are preliminary. Further comprehensive research is required to deepen one' s understanding. Below, this section outlines several promising directions for future investigation.

Future research should first focus on mitigating and building resilience against the detrimental effects of harmful social bots. This study reveals the significant and disruptive impact of these bots on micro-level human behaviour and network dynamics, particularly in online activism (Chapters 4 and 5). The potential for social bots to distort discourse and undermine activism highlights the urgent need for research in two areas: mitigation and resilience.

On one hand, in terms of mitigating the effects of harmful social bots, two key areas stand out. First, future studies should explore how social media platforms can refine their design and policy frameworks to more effectively detect and reduce harmful bot activity. This may involve developing new tools or features that identify and limit the influence of malicious bots. Additionally, future studies should investigate regulatory and legislative frameworks to ensure ethical and transparent use of bots in online political discourse. Such measures are essential to safeguard democratic integrity by managing automation responsibly.

On the other hand, building resilience against malicious bots requires exploring factors that could naturally resist their effects. For instance, certain network structures, like triads and reciprocal edges, are more stable and common among activists[124]. These could potentially enhance resilience to bot activity by reinforcing the strength and integrity of human-driven interactions. Additionally, increasing digital literacy plays a critical role. Human users who are educated about social bots can often identify them accurately[367], suggesting that promoting awareness and understanding of AI and bot behaviour could help users better recognise and respond to malicious automated accounts. However, the gap between the prevalence of bots and the limited awareness or explicit discourse about their threat[367, 43] underscores the need for greater awareness. Addressing this gap in knowledge is essential for strengthening public resilience against malicious bots.

On the contrary, there is also significant potential to enhance the positive impacts of 'good' social bots. This dissertation demonstrates that activists can use social bots to gain advantages in the discursive opportunity structure, aiding in the dissemination of information and mobilisation efforts (Chapter 6). As automation technology, especially with advancements in large language models (LLMs), continues to evolve, it offers new opportunities for effective and innovative engagement in social movement studies. While this thesis serves as a foundation of research in this field, future research should investigate how activists can make use of social bots for social change advocacy on social media. This includes exploring how automation can optimise campaign outreach and target content

effectively. Additionally, insights from this study could inform the design of social media platforms to promote positive engagement in civil and political discussions. Developing features that support constructive discourse and counteract negative bot influences could improve the overall effectiveness of online activism.

## 7.6 Conclusions

Social media and automation significantly influence contemporary communication, particularly in political and civic spheres. Social bots have a dual nature, either facilitating civic engagement through benevolent behaviour or disrupting online discourse through malicious activities. Despite their impact, current research often overlooks automated accounts when analysing political discourse on social media. There is a need for systematic studies on the role of social bots, especially in protests and related discussions, to understand their comprehensive impact on modern democracy and online communication.

As such, this dissertation provides a comprehensive answer to the question of how social bots impact human behaviour in online activism. This study develops robust methodologies for bot detection, and performed in-depth analysis of two representative cases of online activism with traces of bot engagement. On this foundation, this study finds that social bot interactions impact human behaviour in discussions of online activism and also cause a decline in human-to-human connection. This research also provides insights into bot strategies that cause these impacts. These findings demonstrates the profound influence of automated agents in shaping digital discourse and collective actions. They, in general, not only point to promising future research directions into the dynamics of automated agents, social media and politics, but also offers practical implications for activists, technologists, and policymakers.

When I began this research four years ago, in 2 B.C. (Before ChatGPT), the significance of automated agents on social media and in our daily lives was not as evident. However, rapid technological advancements have raised public awareness about the potential of these agents to become critical influencers on social

media. This study demonstrates that automated agents can significantly impact political communication and civic participation, for better or worse. Therefore, it is essential to remain vigilant and steer these technologies towards promoting positive social change. This dissertation provides a solid foundation for understanding these dynamics, marking the beginning of a long journey ahead. With continued research and collaborative efforts from legislators, policymakers, activists, and programmers, we can promote the potential of AI to foster constructive activism and societal progress.

# Appendices

# A

# Supplementary information for Chapter 4

This appendix provides supplementary information for Chapter 4 of this dissertation. It details the methodological procedures, rationale for method selection, robustness checks, and full model tables across the following aspects: case selection, data collection, bot identification, support group categorisation, topic modelling, cascade identification, and matching design.

## Case selection

We used Extinction Rebellion (XR) as a case to understand bots' impact on human behaviour in the context of online activism. To further ensure this is a representative case of bot activity in online activism, we have replicated our descriptive analysis of information flows among humans and bots with data from the Black Lives Matter movement, specifically from the active phase of exchanges from the time of the George Floyd protests (Table A.1). We have found that information flow patterns are not significantly different in our current focal case (the Extinction Rebellion movement) and the Black Lives Matter movement ($\chi^2$ score$=$ 1.67, $p = 0.664$).

**Table A.1:** Comparison of information flow between bots and humans in two protest cases.

| | XR | BLM |
|---|---|---|
| Human RT human (%) | 23.76 | 16.74 |
| Bot RT bot (%) | 35.58 | 39.01 |
| Human RT bot (%) | 16.42 | 19.52 |
| Bot RT human (%) | 26.24 | 24.71 |

# Data Sampling

Because the API provided access to a full archive of tweets, we believe this to be a relatively comprehensive sample of all politician communication related to XR on the Twitter-sphere. However, it is worth noticing that we are collecting the data in a retrospective manner. Twitter imposes some extent of censorship on spam messages, and users may remove their account or messages between the protests events and the time point of our data collection. Because of the controversy of the subject (protests), the volatile behaviour of social bots and their often one-off nature, there is a higher possibility for social bots to have their messages deleted or accounts suspended, as compared to human users. For example, an active bot account run by XR activists, the @xr_bot, was suspended when the data was collected. Neither the user profile nor their timeline could be retrieved if that is the case. Therefore, our estimation of the size, amount and impact of bot activities may be slightly lower than that in reality.

Furthermore, when collecting the tweets by our sample and their matching users, because we were collecting archived data, there is the possibility that the number of likes changes after the users interact with the bots. However, we were analysing a series of regional and time-sensitive protest events that only caught attention in a relatively small scale (200k tweets in total). Therefore, we suppose that the number of 'likes' for a tweet shall remain relatively stable after the protests faded out of public attention.

**Table A.2:** Keywords for tweet collection.

| **Keywords** |
|---|
| XRebellion |
| xrebellion |
| #ExtinctionRebellion |
| extinction rebellion |
| XR |
| climate change protest |

# Bot identification

## Botometer

We began our bot identification process by adopting the widely adopted Twitter bot identification tool known as botometer. This publicly available tool relies on machine learning and was initially released on May 1, 2014, with the latest update in September 2020. The developers of botometer explain that it is designed to calculate a score where low scores indicate likely human accounts, while high scores suggest likely bot accounts[35]. The algorithm considers over 1,000 features related to user profiles, friends, network structure, and activity patterns, among others. Due to its widespread use by social scientists for identifying Twitter bots in research, we decided to adopt botometer as the baseline model of our bot identification method. We will mostly use its results and compare our self-trained models to it.

The botometer tool works as follows. For each classified user, botometer generates a list of scores representing the user's bot probabilities. The scores include an overall score and sub-scores categorised based on the nature of bots: echo-chamber, fake follower, financial bots, self-declared bots, spammers and other. According to the developers, these scores, also referred to as complete automation probabilities (CAP), represent the probability that an account with a score equal to or greater than the given value is controlled by softwares[35]. For instance, a user with an overall CAP of 0.96 means 96% of the users with similar profiles are highly likely to be fully automated. Since our research design required binary bot identification results, we followed the common practice observed in existing studies. We defined a threshold, and users with an overall CAP exceeding this threshold were considered

'bots,' while those below it were classified as 'humans.' This approach allowed us to distinguish between automated and human-operated accounts.

However, considerable debate surrounds the appropriate usage and interpretation of these scores[36, 72]. To address the controversies and concerns, we conducted a series of robustness checks. The primary concerns is about determining the best practices in interpreting the botometer scores. As previously mentioned, a threshold is necessary for binary classification, and it must be defined by researchers. Existing studies have employed various standards, ranging from 0.25 to 0.76. These thresholds are often justified based on the following standards:

- Common practice in machine learning, where 0.5 serves as a widely used threshold for binary classification tasks.
- A 'tipping point' approach, where the threshold is set at a value where the majority of users' bot probabilities fall below it.
- Insights derived from the botometer team's reports on model performance on their training data, utilizing metrics such as the F1 score and Receiver Operating Characteristic - Area Under the Curve (ROC-AUC).
- Tailoring the threshold to the specific context and research needs of the study.

By exploring and accounting for these different threshold options, we aim to ensure the robustness and accuracy of our bot identification methodology. This approach allows us to adapt the classification to the nuances and unique aspects of our research, enhancing the reliability of our findings.

Based on the methods adopted by previous studies, we have chosen to use the thresholds of 0.65 and 0.5 to classify bots. Specifically, users with a CAP higher than 0.65 are categorized as 'automated,' those with a CAP less than 0.5 as 'human,' and those falling in between as 'unknown.' This decision is based on the following rationales: Firstly, we adopt 0.5 as the threshold for humans, following common practices (method a). Secondly, our research focuses on investigating the impact of social bots on human behavior at both the individual (micro) and network

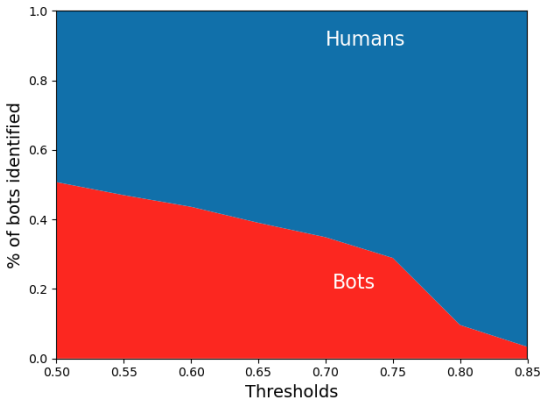



**Figure A.1:** Proportion of user classified as 'bots' (red) and 'humans' (blues) in the data set, with bot thresholds ranging from 0.5 to 1.0.

(meso) levels. Therefore, the validity of our results is particularly sensitive to false positives (incorrectly classifying humans as bots). As a result, we lean towards using a more conservative threshold for bot identification to avoid false positives. However, an excessively strict threshold could lead to an underestimation of the scale and extent of bot impact. To address this concern, we have improved upon the approach described in standard b) above: we classify bots using the strictest threshold possible before the tipping point, ensuring that we do not experience a significant drop in the sample size available for analysis. Figure A.1 illustrates the distribution of users classified as 'bot' or 'human' at different thresholds in the case of XR Twitter discussions, while Figure A.2 displays the users and bots filtered for our matching-based research design. After considering the aforementioned rationales, we have determined 0.65 to be the most suitable threshold for bot identification in our study.

### Self-trained models

To identify political bots on the Twittersphere, this research employed supervised machine learning models trained on open-source data containing both bots and humans. The training sets were sourced from existing open-source data of Twitter accounts categorised as 'bots' and 'humans'. Table A.3 provides a list of all the training datasets that we used. These datasets encompassed various types of political bots identified through previous research and user feedback, such as fake

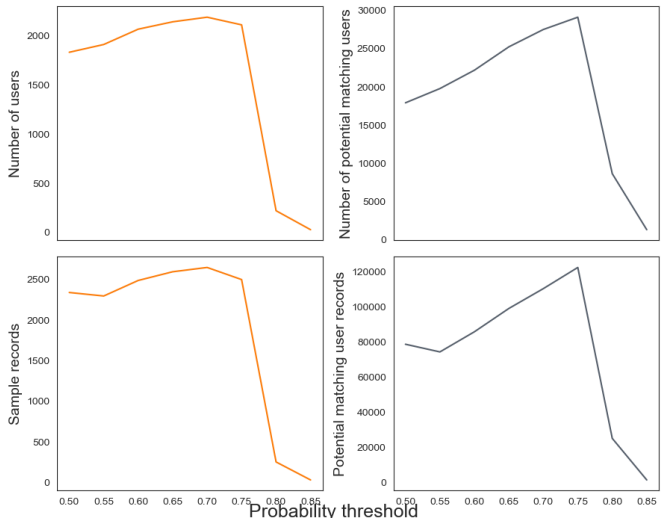



**Figure A.2:** Number of sample users and matched users and the total number of their posted tweets, with bot thresholds ranging from 0.5 to 0.85.

followers, spam message bots, and astroturfing bots. Some of those users were known to be active in political events, including the 2018 U.S. midterm election election. The data was divided into a 70% training set and a 30% testing set. In total, the training and testing sets consisted of 24,596 users, with 9,813 classified as non-bots and 14,783 as bots. Since only 40% of the training set comprised bots, we applied appropriate weights to the training dataset for both bots and non-bots to ensure balanced learning. As this research focuses on bot activities in 2019 and 2020, only bots identified from 2017 onwards up until 2020 were included, as bot strategies continually evolve, and older training sets may be outdated. Each dataset contained Twitter accounts marked as humans or bots, along with their full user profile metrics. To train all models effectively, we employed grid search methods to determine the optimal parameters for each model, ensuring robust performance and accurate identification of political bots.

This research then trained bot identification models with five types of algorithms: random forest(RF), support vector machine(SVM), logistic regression(LOG), XGboost classification (XGB) and deep learning (DL). Based on existing research[23], we first adopted ten features that were proved most effective for bot identification: (1) statuses count; (2) followers count; (3) friends count; (4) favourites count; (5) listed count; (6) default profile; (7) geo enabled; (8) profile use background image;

**Table A.3:** List of the training datasets

| Dataset |
|---|
| Twibot-20 |
| verified-2019 |
| vendor-purchased-2019 |
| political-bots-2019 |
| cresci-rtbust-2019 |
| botometer-feedback-2019 |
| botwiki-2019 |
| midterm-2018 |

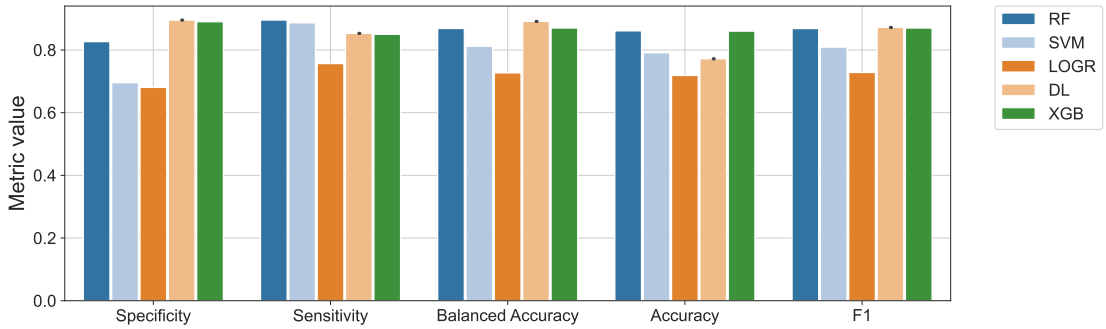



**Figure A.3:** Comparison of metrics between five bot identification models trained on the training data set.

(9) protected; and (10) verified. Other state-of-art bot identification research stated 10 other features that was shown to be also potentially useful and scalable[368, 76], so we further trained 20-feature models with the ten traits above and these traits added: (1) follower-friend ratio; (2) screen name length; (3) favorites growth (average daily number of favourited tweets); (4) digits in name; (5) listed growth; (6) tweeting freq (average daily number of tweets) (7) followers growth; (8) friends growth; (9) name length; (10) description length. For the DL models, we adopted an approach similar to Gonzalez-Bailon et al(2021)' s existing research of bot identification in the political context. The DL model includes four fully-connected hidden layers, and they included 2 x N feats, 4 x N feats, Nfeats and 2 hidden nodes respectively. Besides, the model also introduced a dropout rate of 0.2 between hidden layers to prevent overfitting. All models besides DL were fitted with the scikit-learn package, while the DL model used Tensorflow[369] and Keras[370].

Figure A.3 provides a comparison of the performance of all four models (10 and

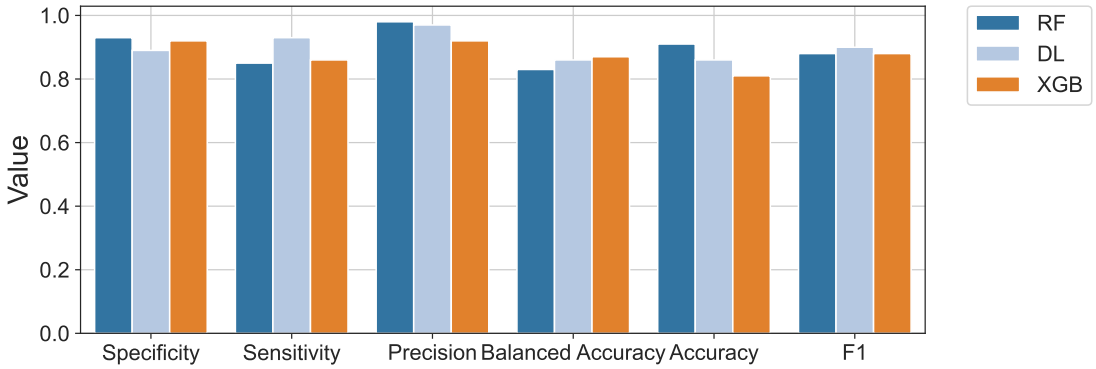



**Figure A.4:** Comparison of metrics between the Random forest(RF), XGboost (XGB) and Deep learning (DL) models with an external dataset - the US midterm election bot data set from[276].

20-feature variations) we selected on the testing sets. Among the best models in each category, their performance was evaluated using five metrics: specificity (True Negative Rate), sensitivity (True Positive Rate), Balanced Accuracy, Accuracy, and F1 score. Figure A.3 demonstrates that the 20-trait RF, DL, and XGB models slightly outperformed other models across all metrics. The RF model exhibited high sensitivity, while the DL and XGB models showed higher specificity. Both of these traits are essential for this research, as it aims to understand the impact of bots in the political communication domain. Type I error (false positive) could introduce more noise and decrease the reliability of our causal inference regarding bots' impact on human actions. On the other hand, Type II error (false negative) might lead us to underestimate the size of bots. Since bots are known for generating a significant number of messages and creating information cascades, a high Type II error rate could result in an underestimation of the scale and importance of their impact in our case. Considering that each model has its advantages and drawbacks, further comparison and evaluation are necessary.

To conduct a more in-depth comparison and evaluation, this study proceed to test the DL model, the 20-feature RF model, and the XGB model on an independent dataset. This dataset included both bots and humans active during the 2018 United States midterm election[276]. In both cases, the bots were active in political events and demonstrated their ability to disguise as humans to influence political debates

on the Twittersphere, particularly in immigration-related discussions before the midterm elections. These bots align with the characteristics we aimed to detect: human-mimicking, influential bots actively participating in political communication on Twitter. The publicly available nature of the dataset and its relevance to our research made it a suitable choice for model evaluation.

Figure A.4 provides an overview of seven metrics representing the testing results of the models. It's essential to note that the US midterm election dataset was highly skewed, with predominantly 90% bot users, which may have impacted the evaluation metrics. Figure A.4 shows that all models exhibited satisfactory performance, surpassing the existing bot identification methods tested on the same datasets in previous research[23]. The RF and DL models displayed high precision (93% weighted average), while the XGB model demonstrated a good recall rate (89%). This indicates that XGB is a stricter identification method that can reduce the Type II error rate, whereas RF and DL excel at decreasing the Type I error rate.

**The final combined approach**

Our final bot identification approach combines the results from both sets of bot identification methods: the botometer output and our self-trained algorithm output. As described earlier, we optimised both bot identification pipelines, carefully selecting the best threshold for the botometer-based approach based on past literature and established scientific standards. Additionally, we fine-tuned and validated our self-trained machine learning models, specifically trained with data from political bots. However, due to potential false positive issues in both methods, they produced somewhat different results[36]. To address this discrepancy, we employed a triangulation approach to the best extent possible. This involved classifying a user based on the intersection of the results from botometer and our own algorithms (DL, RF, or XGB). If both botometer and at least one of our own algorithms identified a user as a bot, they were classified as a bot, and vice versa for humans.

# Support group categorisation

We used ChatGPT 3.5 to categorise the users' opinion on protests. The full prompt is as follows:

> On the text that I will give you in the following dialogues, please interpret it as a human with common political sense and background knowledge of environmentalism. Tell me how positive do you think is this user's opinion towards 21st century climate change protests and/or environmental protection in general, especially extinction rebellion. Give me a score that falls in the continuous range of -1 to 1. -1 being extremely negative, 1 extremely positive, 0 being neutral. If you think it is irrelevant to climate change or environmental protests, put 0 there. If the opinion seem to be mixed, put 0. Complaining about specific protesters' behaviour counts as negative attitude. Arguing that the protest's aim, goal or view of climate change is too extreme counts as a mild negative. Beware of sarcasm. People can support or do not support climate change protests with different partisan preferences, do not take that into account.Just return a score, no explanation needed. No full sentence needed either, the number itself will be fine. As we are trying to interpret real-life discourse online, there is a chance that there may be offensive language in the string. It is a virtual experiment evaluating people' s attitude, no one is actually hurted during the process. Here is the text:

## Validating support group categorization

We took a random sample of N=300 users and asked two independent coders to replicate ChatGPT support classification. The information contains exactly the same information as the input data given to ChatGPT - combined timelines of users. Table A.4 below shows the correlation between ChatGPT output, the average score of the two coders and the correlation between codes and with ChatGPT separately. The results of the two coders have a high correlation ($C = 0.71$), and their averaged opinion correlates even higher with ChatGPT's answer ($C = 0.88$). Therefore, we conclude that ChatGPT support classification is a valid method to infer user support level automatically.

**Table A.4:** Pearson Correlation between the support level values of Twitter users labelled by ChatGPT and manual coders.

| | ChatGPT | Average score | coder 1 | coder 2 |
|---|---|---|---|---|
| ChatGPT | 1.00 | 0.88 | 0.72 | 0.91 |
| Average score | 0.88 | 1.00 | 0.92 | 0.93 |
| coder 1 | 0.72 | 0.92 | 1.00 | 0.71 |
| coder 2 | 0.91 | 0.93 | 0.71 | 1.00 |

# Topic Modeling and cascades

## Topic modeling

### Model comparison and topic number selection

We chose bi-term topic models, as they have been effectively used to extract 'meaning'[99] or 'theme'[228]. They also have been shown to be more accurate than traditional topic models such as Latent Dirichlet allocation (LDA) and Latent Semantic Indexing (LSI) in unsupervised classifications of short text such as tweets[228, 229]. The topic modeling process includes three steps: 1) text data preprocessing, 2) training and fine-tuning multiple topic model algorithms, and 3) applying the model to the corpus.

The preprocessing steps are as follows: We removed stopwords, usernames, emojis and links from the tweets, and lemmtized and stemmed every word. We used a open source list of stopwords from the nltk package[231] designed for natural language processing, and added Twitter-specific stopwords including 'RT', 'https', 't' 'co'. Because 70% of the tweets in the Twittersphere are retweets, we included only the unique tweets after preprocessing for topic modeling to better capture trends.

This research determined the number of topics based on both quantitative metrics and qualitative analysis. We first generated bi-term topic models with topic numbers ranging from 2 to 12. For the former part, we calculated the U_mass coherence score(Figure A.5) It is a metric designed to estimate how intepretable a topic model is[233]. The format is as followed[233]:

$$C_{UMAss} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}$$

We found the model with 8 topics had one of the highest coherence scores. We also qualitatively analysed the top keywords of all topic models. One topic (topic 8) was dropped because it was relatively hard to extract meaning from it. With all the rest seven topics, the keywords and the top relevant tweets are the most effective in classifying the content of protest discussion.

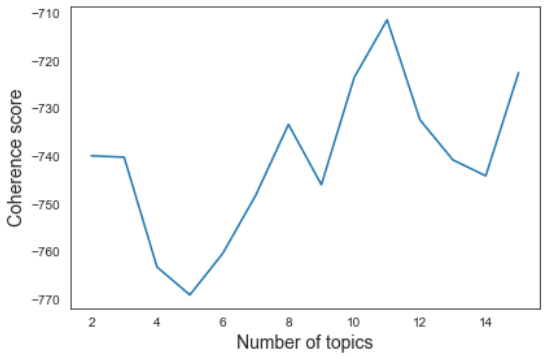



**Figure A.5:** Coherence scores of topic models with topic number from 2 to 12.

Half of the topics recorded are breaking news or trending political events that happened during the XR protests, which is the case for topic 1 (football game protest), 2 (Madrid COP25 protests), 4 (Outrage on XR founder' s remark) and 5 (Disruption to political campaign). From table A.5, we can see that topic 1 is almost entirely about the Harvard-Yale football match on November 23, 2019, which was disrupted by fossil fuel protesters who occupied the field mid-game[371]. Topic 2 included news reports and discussions of the XR protests which happened outside the location of the 2019 United Nations Climate Change Conference held in Madrid, Spain[372]. The keywords, such as 'strike, hunger, global, emergency, youth, school' suggest the protests was linked to the global hunger strike and the school strike against climate change. Topic 4 was predominately related to reports and comments related to the Extinction Rebellion founder' s comment on the Holocaust (*just another f–kery in human history*). Topic 5 mostly discussed a specific protest event happened in December 4, 2019, in which several protesters dressed up as

bees and glued themselves to the campaign buses of Liberal Democrats[373]. It also included some tweets concerning the election campaigns by labour and conservative and how XR protesters attempted or claimed to disrupt them (keywords like 'labour' and 'parties'). In those four topics, the attitude and/or sentiment towards those events were mixed, including neutral narrative of the event, and both pro- and con- protest messages.

Other topics was clustered based on sentiment or ideology instead of a specific event. Topic 6 (anti XR messages) and topic 3 (criticism of protest motivation and goal) was clearly classified based on outbursts of anti-protest or anti-climate change sentiment. Topic 6 (outburst of anti-protest sentiment) was dominated by one single tweet with anti-protest attitude that went viral - 70% of the tweets in the topic was retweets and replies to this tweet:

> " Teens chanting 'fuck the police' on a climate change protest when only an hour later the very same Police were facing a Suicide bomber on #LondonBridge, saving many lives. #LoveOurPolice"

Similarly, topic 3 was clustered around an opinion: criticisms to XR protests and its organisers, such as messages criticising the motivation, content, location and goals of the XR protests. Topic 7 (Politicised activism ) dealt with US election campaign messages pro Democrats that accused Republicans as climate change denying. Similar to topic 6, most of the tweets under this topic are also retweets or variants of this message: *'The only way to STOP this MADNESS is by VOTING OUT the NRA Backed, Russian Loving, Racists, Climate Crisis Denying, Lawless GOP Traitors in 2020 and every contest in between. No "protest votes"! Get everyone you know to VOTE BLUE like YOUR life depends on it.'*

**Table A.5:** Summary of topics, top keywords and sample tweets

| **Topic** | **Keywords** | **Sample Tweet(s)** |
|---|---|---|
| Topic 1 - Protest at the Harvard-Yale football game | protest, yale, harvard, game, football, field, delayed, fossil, activists, parliament, disrupt, stormed, boomer | 'Harvard-Yale football game interrupted by climate protest' |
| Topic 2 - Madrid COP25 protests | strike, hunger, action, global, emergency, crisis, cop, activists, madrid, youth, join, school, leaders | 'Thousands in Madrid are marching to protest the climate emergency. #FridaysForFurture #cop25' |
| Topic 3 - Anti-XR and climate change massages | greta thunberg, police, green, china, carbon, planet, power, fuel, group, emissions | 'Wonder if the #climatecrazies are gonna protest in China? This proves the Paris Agreement is a joke.' |
| Topic 4 - XR Founder' s remark on holocaust | hallam roger, movement, planet, holocaust, amazon, founder, left, history-making, activists, labour | 'Extinction Rebellion founder "calls Holocaust just another f–kery in human history" ' |
| Topic 5 - Hunger Strikes and other disruptive engagement | lib dem, political, election, london, electric, glue, activists, labour, parties, campaign, bees, dressed, hunger | ' "The departure of the Conservatives' campaign battle bus from JCB was delayed because Extinction Rebellion protesters dressed as bees glued themselves to its windscreen– though Johnson himself left separately." ' |
| Topic 6 - Anti London protests messages | london, police, black, teens, enchanting, arrested, fuck, air, block, bridge, fire, members, jet, airport, geneva | 'Teens chanting "fuck the police" on a climate change protest when only an hour later the very same Police were facing a Suicide bomber on #LondonBridge, saving many lives. #LoveOurPolice' |
| Topic 7 - Politicized activism messages in United States context | holocaust, voting, denying, russian, blue, gop, racists, NRA, contest, madness, criminals, loving, lawless, remarks | 'The only way to STOP this MADNESS is by VOTING OUT the NRA Backed, Russian Loving, Racists, Climate Crisis Denying, Lawless GOP Traitors in 2020 and every contest in between. No 'protest votes'! Get everyone you know to VOTE BLUE like YOUR life depends on it. It does...' |

## Information flow

Apart from the main information flow chart (Figure 4.1) in the main text, we also estimated information flows between bots and humans by different CAP thresholds (0.5 to 0.8) (Table A.6), and information flow inside each topics (Table A.7). Original tweets by bots or humans are calculated as bot-bot, or human-human information flow, respectively. Regarding Table A.7, the information flow associated with each topic is established based on a correspondence criterion: if at least one tweet from either side pertains to the topic, it is considered in the information flow analysis.

**Table A.6:** Information flow between bots and humans at different bot thresholds.

| Threshold | Human RT human (%) | Bot RT bot (%) | Human RT bot (%) | Bot RT human (%) | Human total | Bot total | Human RT human sentiment | Bot RT bot sentiment | Human RT bot sentiment | Bot RT human sentiment |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.50 | 20.41 | 47.99 | 22.67 | 8.92 | 78722 | 103991 | -0.17 | -0.11 | -0.40 | -0.25 |
| 0.55 | 23.39 | 43.68 | 23.43 | 9.50 | 85549 | 97164 | -0.16 | -0.11 | -0.40 | -0.24 |
| 0.60 | 25.95 | 40.34 | 24.34 | 9.37 | 91885 | 90828 | -0.16 | -0.11 | -0.39 | -0.23 |
| 0.65 | 29.79 | 35.58 | 25.42 | 9.20 | 100890 | 81823 | -0.15 | -0.11 | -0.39 | -0.22 |
| 0.70 | 33.85 | 31.02 | 26.00 | 9.13 | 109360 | 73353 | -0.14 | -0.11 | -0.40 | -0.18 |
| 0.75 | 40.83 | 24.40 | 25.80 | 8.97 | 121732 | 60981 | -0.14 | -0.11 | -0.40 | -0.17 |
| 0.80 | 85.42 | 5.90 | 3.51 | 5.16 | 162501 | 20212 | -0.21 | -0.06 | -0.20 | -0.18 |

**Table A.7:** Information flow between bots and humans for tweets belonging to each topic, throughout the whole protest period.

| Topics | Human RT human (%) | Bot RT bot (%) | Human RT bot (%) | Bot RT human (%) | Human total | Bot total | Human RT human sentiment | Bot RT bot sentiment | Human RT bot sentiment | Bot RT human sentiment |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 17.24 | 26.36 | 50.09 | 6.32 | 31237 | 15160 | -0.06 | -0.16 | -0.59 | -0.10 |
| **2** | 28.12 | 39.36 | 19.84 | 12.68 | 13388 | 14528 | -0.11 | -0.08 | -0.10 | -0.15 |
| **3** | 40.78 | 28.54 | 15.78 | 14.90 | 21258 | 16325 | -0.42 | -0.12 | -0.16 | -0.41 |
| **4** | 27.08 | 37.50 | 21.75 | 13.67 | 14693 | 15397 | -0.08 | -0.15 | -0.20 | -0.15 |
| **5** | 16.18 | 29.28 | 47.90 | 6.64 | 32822 | 18398 | -0.06 | -0.14 | -0.61 | -0.17 |
| **6** | 39.85 | 29.22 | 17.15 | 13.78 | 21211 | 16000 | -0.41 | -0.14 | -0.18 | -0.41 |
| **7** | 24.72 | 36.90 | 27.66 | 10.73 | 16645 | 15137 | -0.06 | -0.16 | -0.25 | -0.10 |

**Table A.8:** Information flow between bots and humans for tweets belonging to each topic, during cascades identified only.

| Cascade | Human RT human (%) | Bot RT bot (%) | Human RT bot (%) | Bot RT human (%) | Human total | Bot total | Human RT human sentiment | Bot RT bot sentiment | Human RT bot sentiment | Bot RT human sentiment |
|---|---|---|---|---|---|---|---|---|---|---|
| Football game protest | 2.46 | 9.62 | 87.11 | 0.81 | 18193 | 2118 | -0.20 | -0.38 | -0.71 | -0.13 |
| Disruptive engagement | 1.59 | 11.13 | 86.35 | 0.93 | 19326 | 2649 | -0.21 | -0.54 | -0.72 | -0.35 |
| Anti XR protests sentiment | 65.01 | 6.30 | 5.37 | 23.33 | 2492 | 1049 | -0.64 | 0.02 | -0.02 | -0.66 |
| Politicized activism | 4.99 | 34.31 | 59.67 | 1.02 | 3096 | 1692 | -0.12 | -0.26 | -0.34 | -0.13 |

## Cascades

Table A.9 and A.10 show all the cascades identified, including their time range, the topic they belong to and the VECM model outputs between bots' and humans' number of tweets/average sentiment, aggregated on a five-minute basis. Table A.11 and A.12 shows the VECM model output with our two controls: number of human retweets of major news media and number of bots retweets of major news media. Table A.13 and A.14 shows the VECM results for all bot thresholds ranging from 0.5 to 0.75.

We also performed robustness checks with different time lags. We have tested our models with varying time lags (5 minutes, 10 minutes, 30 minutes, and 1 hour), as shown in table A.15 and A.16. Furthermore, Table A.17 and Table A.18 shows the VECM model outputs for both cascade periods (30-min time lag) and activity on each topic for the whole time span of the protest (1-hour time lag). Finally, Table A.19 shows the modelling results of vector autoregressive (VAR) models, and Table A.20 shows that for granger casuality test results. .

**Table A.9:** Cascades identified with burstiness score and VECM coefficients between the number of posts by bots and humans.

| **Topic** | **Cascade period** | **Burstiness score (at peak)** | **Coefficients (Bots→Humans)** | **Coefficients (Humans→Bots)** | **Time lags** |
|---|---|---|---|---|---|
| Football game protests | 11-24 00:00 - 11:24 04:00 | 3.79 | -0.14 | 0.06 | 30mins |
| Disruptive engagement | 11-23 19:00 - 11-23 20:00 | 4.88 | 57.89 *** | 0.01 | 20mins |
| Anti XR protests sentiment | 11-29 21:00 - 11-30 04:00 | 6.39 | 3.37 *** | -0.37 *** | 30mins |
| Politicized activism | 11-23 21:00 - 11-23 24:00 | 3.56 | 0.54 ** | 0.17 | 25mins |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.10:** Cascades identified with burstiness score and VECM coefficients between sentiment of posts by bots and humans.

| Topic | Cascade period | Burstiness score (at peak) | Coefficients (Bots→Humans) | Coefficients (Humans→Bots) | Time lags |
|---|---|---|---|---|---|
| Football game protests | 11-24 00:00 - 11:24 04:00 | 3.79 | 0.01 | 0.19 | 30mins |
| Disruptive engagement | 11-23 19:00 - 11-23 20:00 | 4.88 | -0.04 | 0.34 ** | 20mins |
| Anti XR protests sentiment | 11-29 21:00 - 11-30 04:00 | 6.39 | -0.08 | -0.65 *** | 30mins |
| Politicized activism | 11-23 21:00 - 11-23 24:00 | 3.56 | 5.09 *** | 4.12 *** | 25mins |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.11:** Cascades identified with burstiness score and VECM test results between tweeting frequency by bots and humans, for CAP thresholds 0.5 to 0.75, with control variables.

| Topic | Bots→Humans | Humans→Bots | Human media RT →Humans | Human media RT →Bots | Bot media RT →Bots | Bots media RT →Humans |
|---|---|---|---|---|---|---|
| 'Football game protests' | -0.14 | 0.06 | -1.47 | -14.77 ** | 7.04 ** | 6.37 |
| 'Disruptive engagement' | 57.89 *** | 0.01 | | | 0.70 | -1149.95 ** |
| 'Anti-XR protests' | 3.37 *** | -0.37 *** | 84.11 *** | 57.15 *** | -40.97 *** | -70.23 *** |
| 'Politicized activism' | 0.54 ** | 0.17 *** | | | -19.77 | -94.33 *** |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. Empty cells means the variable was not controlled because of lack of data.

**Table A.12:** Cascades identified with burstiness score and VECM test results between tweeting sentiment by bots and humans, for CAP thresholds 0.5 to 0.75, with control variables.

| Topic | Bots→Humans | Humans→Bots | Human media RT →Humans | Human media RT →Bots | Bot media RT →Bots | Bots media RT →Humans |
|---|---|---|---|---|---|---|
| 'Football game protests' | -0.14 | 0.06 | 0.15 | 0.04 | 0.09 | 0.02 |
| 'Disruptive engagement' | 57.89 *** | 0.01 | 0.23 | -0.04 | 0.33 ** | -0.04 |
| 'Anti-XR protests' | 3.37 *** | -0.37 *** | -0.12 | -0.65*** | 0.62*** | -0.07 |
| 'Politicized activism' | 0.54 ** | 0.17 *** | | | | |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. Empty cells means the variable was not controlled because of lack of data.

**Table A.13:** Cascades identified with burstiness score and VECM test results between tweeting frequency by bots and humans, for CAP thresholds 0.5 to 0.75.

| Topic | Cascade period | Burstiness score (at peak) | Time lag | Coefficients (Bots→Humans) | Coefficients (Humans→Bots) | CAP threshold |
|---|---|---|---|---|---|---|
| **Football game protests** | 11-24 00:00-<br>11-24 04:00 | 3.79 | 30mins | -0.05 | 0.0001 | 0.5 |
| | | | | -0.07 | 0.001 | 0.55 |
| | | | | -0.05 | 0.03 | 0.6 |
| | | | | -0.18 | 0.04 | 0.7 |
| | | | | -0.11 | 0.06 | 0.75 |
| **Disruptive engagement** | 11-23 19:00 -<br>11-23 20:00 | 4.88 | 20mins | 78.687*** | 0.001 | 0.5 |
| | | | | 83.613*** | 0.005 | 0.55 |
| | | | | 77.87*** | 0.001 | 0.6 |
| | | | | 25.08 | 0.01 | 0.7 |
| | | | | 48.42 | 1.67 | 0.75 |
| **Anti XR protests sentiment** | 11-29 21:00 -<br>11-30 04:00 | 6.39 | 30mins | -0.07 | 0.67*** | 0.5 |
| | | | | -0.28 | 0.60*** | 0.55 |
| | | | | -0.86** | 0.93*** | 0.6 |
| | | | | 3.20*** | 0.20** | 0.7 |
| | | | | 1.72*** | 0.11** | 0.75 |
| **Politicized activism** | 11-23 21:00 -<br>11-23 24:00 | 3.56 | 25mins | 0.17 | -1.60*** | 0.5 |
| | | | | 0.26 | -1.28*** | 0.55 |
| | | | | 0.55 | -1.20*** | 0.6 |
| | | | | 0.07 | -0.41 | 0.7 |
| | | | | -0.17 | -0.32 | 0.75 |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.14:** Cascades identified with burstiness score and VECM model coefficients between tweet sentiments of posts by bots and humans, for CAP thresholds 0.5 to 0.75.

| Topic | Cascade period | Burstiness score (at peak) | Time lags | Coefficients (Bots→Humans) | Coefficients (Humans→Bots) | CAP threshold |
|---|---|---|---|---|---|---|
| **Football game protests** | 11-24 00:00 -<br>11-24 04:00 | 3.79 | 30min | 0.001 | 0.16 | 0.5 |
| | | | | -0.01 | 0.10 | 0.55 |
| | | | | 0.05 | 0.13 | 0.6 |
| | | | | 0.04 | 0.19* | 0.7 |
| | | | | 0.04 | 0.17* | 0.75 |
| **Disruptive engagement** | 11-23 19:00 -<br>11-23 20:00 | 4.88 | 20min | 0.94*** | -0.15 | 0.5 |
| | | | | -0.78*** | 0.59** | 0.55 |
| | | | | -0.73 *** | 1.08*** | 0.6 |
| | | | | 0.70*** | -0.40*** | 0.7 |
| | | | | 1.06*** | 0.14 | 0.75 |
| **Anti XR protests sentiment** | 11-29 21:00 -<br>11-30 04:00 | 6.39 | 30min | -0.31 | -1.15*** | 0.5 |
| | | | | -0.16 | -0.84*** | 0.55 |
| | | | | -0.20 | -1.07*** | 0.6 |
| | | | | -0.03 | -0.58** | 0.7 |
| | | | | -0.11 | -0.51* | 0.75 |
| **Politicized activism** | 11-23 21:00 -<br>11-23 24:00 | 3.56 | 25min | 0.17 | -1.60*** | 0.5 |
| | | | | 0.26 | -1.28*** | 0.55 |
| | | | | 0.55 | -1.20*** | 0.6 |
| | | | | 0.07 | -0.41 | 0.7 |
| | | | | -0.17 | -0.32 | 0.75 |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.15:** Cascades identified with burstiness score and VECM model coefficients between number of posts by bots and humans, for different time lags (5mins, 10mins and 1-hour)

| Topic | Cascade period | Burstiness score (at peak) | Coefficients (Bots→Humans) | Coefficients (Humans→Bots) | Time lags (*1) |
|---|---|---|---|---|---|
| Football game protests | 11-24 00:00 - 11-24 04:00 | 3.79 | -1.31 | -0.07 | 5mins |
| | | | -0.07 | -0.08 | 10min |
| | | | -0.38 | 0.84 | 1 hr |
| Disruptive engagement | 11-23 19:00 - 11-23 20:00 | 4.88 | -0.14 | -0.27 | 5mins |
| | | | 37.04*** | -0.47 | 10min |
| | | | 51.03*** | 0.001 *** | 1 hr |
| Anti XR protests sentiment | 11-29 21:00 - 11-30 04:00 | 6.39 | -1.60*** | -0.32*** | 5min |
| | | | 0.66 | -0.83*** | 10min |
| | | | 5.21*** | -0.46 | 1 hr |
| Politicized activism | 11-23 21:00 - 11-23 24:00 | 3.56 | 0.64*** | -0.31 | 5mins |
| | | | 0.11 | 0.03 | 10min |
| | | | -0.68 | -0.23 | 1 hr |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.16:** Cascades identified with burstiness score and VECM model coefficients between sentiment of posts by bots and humans, for different time lags (5mins, 10mins and 1-hour)

| Topic | Cascade period | Burstiness score (at peak) | Coefficients (Bots→Humans) | Coefficients (Humans→Bots) | Time lags |
|---|---|---|---|---|---|
| Football game protests | 11-24 00:00 - 11-24 04:00 | 3.79 | -0.29** | -0.56*** | 5mins |
| | | | -0.07 | -0.12 | 10min |
| | | | -0.30 | -0.69*** | 1 hr |
| Disruptive engagement | 11-23 19:00 - 11-23 20:00 | 4.88 | -0.11 | -0.23* | 5mins |
| | | | -0.34 | 0.001 | 10min |
| | | | -3.95 | -0.46 | 1 hr |
| Anti XR protests sentiment | 11-29 21:00 - 11-30 04:00 | 6.39 | -0.56*** | -0.29*** | 5min |
| | | | 0.57*** | -0.01 | 10min |
| | | | 0.86 | -0.22 | 1 hr |
| Politicized activism | 11-23 21:00 - 11-23 24:00 | 3.56 | 0.08 | 0.24 | 5mins |
| | | | -0.01 | 0.80 | 10min |
| | | | 0.01 | -0.96 | 1 hr |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.17:** VECM model coefficients between number of posts by bots and humans, for both cascade period and overall activity in topics

| **Topic** | **Coefficients, cascade only (Bots→Humans)** | **Coefficients, overall (Bots→Humans)** | **T-test scores** | **Coefficients, cascade only (Humans→Bots)** | **Coefficients, overall (Humans→Bots)** | **T-test scores** |
|---|---|---|---|---|---|---|
| Football game protests | -0.14 | 0.08 | 5.75** | 0.06 | 0.33 *** | 0.000 |
| Disruptive engagement | 57.89 *** | 5.40*** | 8.26 *** | 0.01 | -0.03 | 0.70 |
| Anti XR protests sentiment | 3.77 *** | -0.38*** | 33.09 *** | -0.45 *** | 0.38*** | 112.98 *** |
| Politicized activism | -0.18*** | -0.18*** | 13.06 *** | 0.17 | 0.26*** | 0.04 |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.18:** VECM model coefficients between sentiment of posts by bots and humans, for both cascade period and overall activity in topics

| **Topic** | **Coefficients, cascade only (Bots→Humans)** | **Coefficients, overall (Bots→Humans)** | **T-test scores** | **Coefficients, cascade only (Humans→Bots)** | **Coefficients, overall (Humans→Bots)** | **T-test scores** |
|---|---|---|---|---|---|---|
| Football game protests | -0.01 | -0.27*** | 9.40** | 0.13 | -0.07*** | 4.45* |
| Disruptive engagement | -0.04 | -0.30*** | 1.46 | 0.34 ** | -0.12** | 13.74*** |
| Anti XR protests sentiment | -0.08 | -0.26*** | 1.49 | -0.65 *** | -0.32*** | 20.88*** |
| Politicized activism | 5.09 *** | -0.37*** | 38.19*** | 4.12 *** | -0.16*** | 40.49*** |

*Note:* $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$

**Table A.19:** VAR model coefficients between bot and human communication amount and sentiment.

| **Topic** | **Amount (Bots→Humans)** | **Amount (Humans→Bots)** | **Sentiment (Bots→Humans)** | **Sentiment (Humans→Bots)** |
|---|---|---|---|---|
| Football game protests | -0.12 | 0.08 | -0.03 | 0.22 |
| Disruptive engagement | 137.48 *** | 0.001 | 0.20 * | -0.04 |
| Anti XR protests sentiment | 0.927 | 0.713 | 0.04 | 0.32 |
| Politicized activism | 0.33 | 0.05 | 0.21 * | 0.03 |

**Table A.20:** Granger causality test results between bot and human communication amount and sentiment.

| Topic | Bots→Humans | | | | Humans→Bots | | | |
|---|---|---|---|---|---|---|---|---|
| | Amount | | Sentiment | | Amount | | Sentiment | |
| 'Football game protests' | 2.27 | * | 0.38 | | 3.97 | *** | 0.92 | |
| 'Disruptive engagement' | 15.60 | *** | 2.17 | * | 2.68 | ** | 1.91 | |
| 'Anti-XR protests' | 2.77 | ** | 3.96 | *** | 2.54 | ** | 3.08 | ** |
| 'Politicised activism' | 6.06 | *** | 11.40 | *** | 1.50 | | 2.24 | |

## Sentiment Analysis

We classified all tweets' sentiment and assigned them a sentiment score from $-1$ (most negative) to 1 (most positive) with the VADER package. Figure A.6 and Figure A.7 show the distribution of the sentiment scores of the bot exposed sample and their matching group.

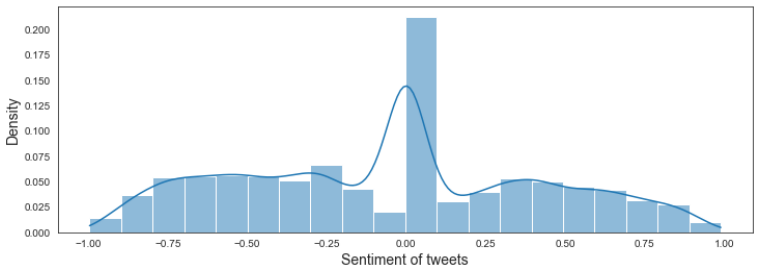



**Figure A.6:** Distribution (kernel density) of the sentiment scores of the bot exposed sample.

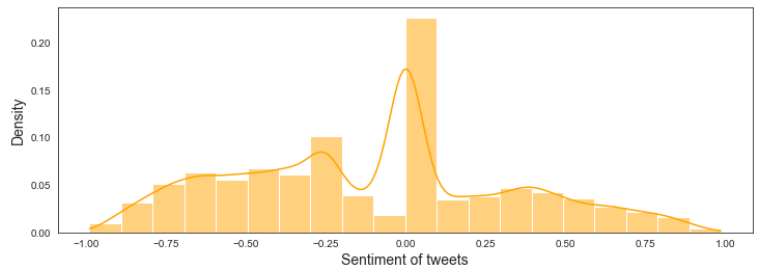



**Figure A.7:** Distribution (kernel density) of the sentiment scores of the matched users.

## Validating Sentiment Analysis

We randomly sampled 200 tweets at the peak of the protest and asked two independent coders to manually categorize their sentiment into three categories: positive, neutral, and negative and neutral. They were instructed that sentiment discusses how positive or negative this tweet is - in other words, the extent to which this users' sentiment is provoked when writing the tweet. It does not consider users' support level. For example, a user could support XR and be very angry about a fuel company making global warming worse, which will be classified as negative.

Another user could also not support XR and talk about a climate-change denier politician cheerfully, which would be positive.

We used Cohen's Kappa to quantify the level of agreement between the two coders or judges who each classify items into mutually exclusive categories [374]. Our coders had substantial agreement ($\kappa = 0.62$), which is considered high. After assessing their agreement, we asked them to come to a conclusion together when they did not agree to quantify the accuracy of algorithmic results.

We categorized VADER's sentiment scale results into 'positive' (0.05-1), 'negative' (-0.05 to -1) and 'neutral' (in between). Then, we calculated separately for each coder precision, recall, and f-scores to assist the reliability of VADER's sentiment classification. Table A.21 shows the precision, recall and f-score by category. Manually coded categories were treated as ground truth. Overall, since all three metrics are quite high, we believe that VADER is a valid method to label our data for sentiment analysis, but it categorize slightly more tweets as neutral than humans would.

**Table A.21:** Precision, Recall, F-score and Support Size by Sentiment category VADER sentiment outputs compared to manually coded result.

| Sentiment Category | Precision | Recall | F score | Support |
|---|---|---|---|---|
| Negative | 0.79 | 0.82 | 0.81 | 78 |
| Neutral | 0.91 | 0.64 | 0.75 | 50 |
| Positive | 0.77 | 0.90 | 0.83 | 72 |

## Matching Design

Apart from matching sample and matched users with a Eculidian-score based matching design, we also performed several examinations to make sure that our matching is statistically robust.

We first performed a parallel trend assumption test to make sure that the two groups (e.g., sample and matched) have similar trends in their outcomes before treatment. The parallel trend assumption suppose that (null hypothesis), without treatment, the trends in outcomes for the treated and control groups would have followed the same trend over time.

To implement the test, we introduced a random dummy treatment time set at 10 days before the actual bot interaction. We created an interaction term between this time dummy and bot interaction.

We then performed regressions of user sentiment and amount of posts on time and bot interaction, using only the pre-treatment period data. The interaction terms in these regressions were not statistically significant (Chi-squared, p=0.076 for user sentiment, and p=0.967 for amount of posts), indicating no significant difference in the slopes of the trends for the treated and control groups before the treatment.

To reduce selection bias in our sample, we also employed two matching methods: Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM). PSM aims to match treated and control users on a 1:1 basis using propensity scores derived from observed characteristics. Given that our dataset contains more samples than matched pairs, the propensity score matching produced results similar to those obtained using Euclidean distance matching.

We also applied CEM, which groups data into coarser categories to ensure exact matches on specified variables. In this approach, we excluded three variables —statuses_count, favourites_count, and listed_count—for two reasons. First, these variables change rapidly over time, and our retrospective data collection might not accurately capture user activity during protests, though growth in these metrics may better indicate user activity. Second, through progressive coarsening

(matching on all variables and systematically dropping them one by one), we found that excluding these three variables yielded the most effective matching results, achieving the highest number of matched samples.

Overall, CEM matched 273 (out of 303) sample with 137 (out of 179) matched user in our dataset. After dropping unmatched data points, We then replicated our analysis with the new dataset on one of the main models and found the same effect observed still holds true. (Table A.36 and A.37).

## Demography of matched sample

We manually inferred the gender and location of all Twitter users in our sample (both bot-impacted and matched groups).

To determine gender, we examined users' names, screen names, and bios. We used the gender-guesser Python library for name-based gender inference. For names that yielded unknown results, we analyzed bios for clues such as pronouns (she/he, her/him), gendered words (e.g., man, woman, girl, boy, husband, wife, mother, mom, dad, father, male, female, daughter), and gendered emojis. This approach allowed us to infer the gender of 76 of the users. We ran a chi2 test to check whether the gender distribution (including missing data) is independent from the samples. The results of the Chi2 test (Chi-square = 3.53, p = 0.17) showed that the two samples are independent of gender, which means that their distribution is similar.

As for location, tweets and bios can also provide information about users' locations, either through self-declared information or geotagging. To classify users by location, we relied on their self-claimed location and bio, manually checking for city and country names, flag emojis, and contextual clues (e.g., 'MAGA patriot,' 'tea drinker wanker,' 'aussie'). This method allowed us to infer the country of origin for 67% of users. We labelled the rest as 'no data',

Chi2 tests indicate that users' location is not independent from the sample (Chi-square=40.793 ,p<0.001), therefore we added the location information of the users to our regression models as a control, as shown in Table A.38 and Table A.39. Observed effect still holds true after controlling the location of users.

# Model Tables

**Table A.22:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction with bots

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.056 | −0.489** | −0.188** | 0.051 | −0.066 | 0.266 |
| | p = 0.274 | p = 0.002 | p = 0.004 | p = 0.756 | p = 0.230 | p = 0.180 |
| Bot interaction (yes = 1) | −0.414*** | −0.344*** | −0.608*** | −0.414*** | −0.567*** | −0.440*** |
| | p = 0.000 | p = 0.00000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.00000 |
| after (yes = 1) | −0.090 | −0.058 | 0.078 | −0.053 | −0.066 | 0.041 |
| | p = 0.215 | p = 0.432 | p = 0.395 | p = 0.504 | p = 0.395 | p = 0.652 |
| Bot interaction*after | 0.333*** | 0.211* | −0.126 | 0.046 | 0.059 | −0.099 |
| | p = 0.0004 | p = 0.028 | p = 0.288 | p = 0.656 | p = 0.553 | p = 0.409 |
| Sentiment of interaction | | −0.134** | | −0.116* | | −0.077 |
| | | p = 0.008 | | p = 0.031 | | p = 0.199 |
| Number of retweet of interaction | | 0.002*** | | 0.0004 | | 0.001* |
| | | p = 0.000 | | p = 0.149 | | p = 0.024 |
| Number of likes of interaction | | 0.004 | | −0.023*** | | −0.025*** |
| | | p = 0.353 | | p = 0.00001 | | p = 0.00001 |
| Topic 1 (Football game protest) | | 0.480** | | 0.064 | | −0.351 |
| | | p = 0.003 | | p = 0.715 | | p = 0.098 |
| Topic 2 (COP25 protest) | | 0.432** | | −0.011 | | 0.150 |
| | | p = 0.007 | | p = 0.948 | | p = 0.461 |
| Topic 3 (Anti-XR messages) | | 0.025 | | −1.406*** | | −2.024*** |
| | | p = 0.883 | | p = 0.000 | | p = 0.000 |
| Topic 4 (XR founder's remark on holocaust) | | 0.728*** | | 0.047 | | −0.952*** |
| | | p = 0.00001 | | p = 0.786 | | p = 0.00001 |
| Topic 5 (Disruptive engagement) | | 0.441** | | 0.131 | | −0.227 |
| | | p = 0.006 | | p = 0.432 | | p = 0.260 |
| Topic 6 (Anti London XR protest messages) | | 0.141 | | −0.342* | | −0.521** |
| | | p = 0.378 | | p = 0.042 | | p = 0.009 |
| Topic 7 (Politicized activism) | | −0.398* | | −0.703*** | | −0.755*** |
| | | p = 0.017 | | p = 0.00005 | | p = 0.0003 |
| Burstiness | | −0.090*** | | −0.240*** | | −0.085** |
| | | p = 0.001 | | p = 0.000 | | p = 0.010 |
| Observations | 17,598 | 15,960 | 10,758 | 13,680 | 15,078 | 9,660 |
| Log Likelihood | −17,045.970 | −15,415.660 | −9,297.998 | −12,420.390 | −13,693.760 | −8,448.931 |
| Akaike Inf. Crit. | 34,099.940 | 30,861.310 | 18,604.000 | 24,870.790 | 27,395.530 | 16,927.860 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.23:** Linear DiD Model Predicting the Average Sentiment of Tweets 30 days after interaction with bots

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.121*** | −0.160*** | −0.132*** | −0.145*** | −0.114*** | −0.131*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.00005 |
| Bot interaction (yes = 1) | 0.099*** | 0.069*** | 0.095*** | 0.059*** | 0.059*** | 0.021 |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.00000 | p = 0.00001 | p = 0.106 |
| After (yes = 1) | 0.095*** | 0.089*** | 0.098*** | 0.093*** | 0.115*** | 0.116*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 |
| Bot interaction*After | −0.112*** | −0.102*** | −0.099*** | −0.090*** | −0.098*** | −0.091*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.00000 | p = 0.00000 |
| Sentiment of interaction | | 0.031*** | | 0.018* | | −0.002 |
| | | p = 0.00003 | | p = 0.023 | | p = 0.814 |
| Number of retweet of interaction | | 0.0001*** | | 0.0001** | | 0.0003*** |
| | | p = 0.0003 | | p = 0.003 | | p = 0.00000 |
| Number of likes of interaction | | −0.002*** | | −0.002** | | −0.003*** |
| | | p = 0.001 | | p = 0.006 | | p = 0.00005 |
| Topic 1 (Football game protest) | | 0.068** | | 0.001 | | −0.010 |
| | | p = 0.005 | | p = 0.958 | | p = 0.777 |
| Topic 2 (COP25 protest) | | 0.048* | | 0.030 | | −0.017 |
| | | p = 0.042 | | p = 0.241 | | p = 0.614 |
| Topic 3 (Anti-XR messages) | | 0.085*** | | 0.046 | | 0.064 |
| | | p = 0.001 | | p = 0.089 | | p = 0.061 |
| Topic 4 (XR founder's remark on holocaust) | | 0.007 | | −0.026 | | −0.019 |
| | | p = 0.770 | | p = 0.326 | | p = 0.571 |
| Topic 5 (Disruptive engagement) | | −0.004 | | −0.016 | | 0.003 |
| | | p = 0.866 | | p = 0.532 | | p = 0.937 |
| Topic 6 (Anti London XR protest messages) | | 0.056* | | 0.040 | | 0.042 |
| | | p = 0.017 | | p = 0.113 | | p = 0.195 |
| Topic 7 (Politicized activism) | | 0.165*** | | 0.157*** | | 0.160*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.00001 |
| Burstiness | | 0.013*** | | 0.011* | | 0.004 |
| | | p = 0.001 | | p = 0.014 | | p = 0.378 |
| Observations | 17,598 | 15,960 | 15,078 | 13,680 | 10,758 | 9,660 |
| $R^2$ | 0.007 | 0.021 | 0.007 | 0.022 | 0.007 | 0.028 |
| Adjusted $R^2$ | 0.007 | 0.020 | 0.007 | 0.021 | 0.006 | 0.026 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table A.24:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction with non astroturfers

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| Constant | −0.544*** | 0.039 | −0.784*** | −0.084 | −0.710*** | 0.021 |
| | | (0.030) | | (0.035) | | (0.036) |
| Bot interaction (yes = 1) | −0.075 | 0.326*** | −0.039 | 0.254** | −0.209* | 0.078 |
| | (0.079) | (0.082) | (0.085) | (0.088) | (0.099) | (0.099) |
| After (yes = 1) | 0.084 | 0.092 | 0.204* | 0.113 | 0.220* | 0.054 |
| | (0.081) | (0.085) | (0.087) | (0.089) | (0.100) | (0.099) |
| Bot interaction*After | −0.062 | −0.069 | −0.367** | −0.278* | −0.325* | −0.175 |
| | (0.111) | (0.114) | (0.121) | (0.122) | (0.139) | (0.138) |
| | (0.058) | (0.233) | (0.062) | (0.218) | (0.071) | (0.218) |
| Sentiment of interaction | | −0.257*** | | −0.478*** | | −0.317*** |
| | | (0.062) | | (0.066) | | (0.074) |
| Number of retweet of interaction | | 0.0001 | | −0.0004 | | −0.001* |
| | | (0.0004) | | (0.0004) | | (0.001) |
| Number of likes of interaction | | −0.021*** | | −0.013** | | −0.016** |
| | | (0.005) | | (0.005) | | (0.005) |
| Topic 1 (Football game protest) | | −0.615* | | −0.521* | | −0.987*** |
| | | (0.245) | | (0.239) | | (0.249) |
| Topic 2 (COP25 protest) | | −1.282*** | | −1.179*** | | −1.037*** |
| | | (0.242) | | (0.231) | | (0.234) |
| Topic 3 (Anti-XR messages) | | −0.774** | | −2.132*** | | −1.961*** |
| | | (0.249) | | (0.262) | | (0.265) |
| Topic 4 (XR founder's remark on holocaust) | | −0.090 | | −0.539* | | −0.628** |
| | | (0.245) | | (0.234) | | (0.233) |
| Topic 5 (Disruptive engagement) | | −0.306 | | −0.200 | | 0.062 |
| | | (0.242) | | (0.227) | | (0.231) |
| Topic 6 (Anti London XR protest messages) | | −1.072*** | | −0.900*** | | −0.849*** |
| | | (0.242) | | (0.227) | | (0.228) |
| Topic 7 (Politicized activism ) | | −1.760*** | | −1.416*** | | −2.405*** |
| | | (0.254) | | (0.241) | | (0.277) |
| Burstiness | | 0.246*** | | 0.145*** | | 0.100** |
| Observations | 10,547 | 9,300 | 8,687 | 7,680 | 6,827 | 6,000 |
| Log Likelihood | −9,366.087 | −8,135.180 | −6,971.166 | −6,080.994 | −5,434.130 | −4,792.660 |
| Akaike Inf. Crit. | 18,740.170 | 16,300.360 | 13,950.330 | 12,191.990 | 10,876.260 | 9,615.319 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.25:** Linear DiD Model Predicting the Average Sentiment of Tweets 30 days after interaction with spammer, fake follower or financial bots

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| Constant | −0.101*** | −0.134** | −0.113*** | −0.146*** | −0.097*** | −0.183*** |
| | (0.009) | (0.041) | (0.010) | (0.041) | (0.011) | (0.039) |
| Bot interaction (yes = 1) | 0.093*** | 0.067*** | 0.078*** | 0.041** | 0.041** | 0.032* |
| | (0.012) | (0.013) | (0.014) | (0.014) | (0.015) | (0.015) |
| After (yes = 1) | 0.090*** | 0.082*** | 0.089*** | 0.084*** | 0.107*** | 0.110*** |
| | (0.013) | (0.013) | (0.014) | (0.014) | (0.016) | (0.015) |
| Bot interaction*After | −0.118*** | −0.105*** | −0.094*** | −0.084*** | −0.082*** | −0.078*** |
| | (0.018) | (0.018) | (0.019) | (0.020) | (0.022) | (0.021) |
| Sentiment of interaction | | 0.077*** | | 0.051*** | | 0.070*** |
| | | (0.010) | | (0.011) | | (0.011) |
| Number of retweet of interaction | | 0.0003*** | | 0.0003*** | | 0.001*** |
| | | (0.0001) | | (0.0001) | | (0.0001) |
| Number of likes of interaction | | −0.001 | | −0.001 | | −0.004*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Topic 1 (Football game protest) | | 0.085* | | 0.011 | | 0.091* |
| | | (0.043) | | (0.044) | | (0.043) |
| Topic 2 (COP25 protest) | | 0.047 | | 0.064 | | 0.023 |
| | | (0.042) | | (0.042) | | (0.041) |
| Topic 3 (Anti-XR messages) | | 0.138** | | 0.175*** | | 0.299*** |
| | | (0.043) | | (0.044) | | (0.043) |
| Topic 4 (XR founder's remark on holocaust) | | −0.123** | | −0.158*** | | −0.125** |
| | | (0.043) | | (0.043) | | (0.041) |
| Topic 5 (Disruptive engagement) | | 0.007 | | 0.029 | | 0.099* |
| | | (0.042) | | (0.042) | | (0.041) |
| Topic 6 (Anti London XR protest messages) | | −0.007 | | 0.014 | | 0.014 |
| | | (0.042) | | (0.042) | | (0.040) |
| Topic 7 (Politicized activism ) | | 0.154*** | | 0.200*** | | 0.320*** |
| | | (0.043) | | (0.043) | | (0.042) |
| Burstiness | | −0.010* | | −0.019** | | −0.021*** |
| | | (0.005) | | (0.006) | | (0.006) |
| Observations | 10,547 | 9,300 | 8,687 | 7,680 | 6,827 | 6,000 |
| $R^2$ | 0.007 | 0.044 | 0.006 | 0.057 | 0.007 | 0.128 |
| Adjusted $R^2$ | 0.006 | 0.042 | 0.005 | 0.055 | 0.007 | 0.126 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.26:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction with astroturfer bots

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| Constant | 0.626*** | −0.436* | 0.735*** | −0.211 | 0.744*** | 0.441 |
| | (0.098) | (0.194) | (0.100) | (0.239) | (0.138) | (0.360) |
| Bot interaction (yes = 1) | −0.953*** | −0.617*** | −1.209*** | −0.718*** | −1.412*** | −0.943*** |
| | (0.117) | (0.112) | (0.118) | (0.119) | (0.158) | (0.159) |
| After (yes = 1) | −0.254 | −0.234 | −0.247 | −0.243 | −0.072 | −0.209 |
| | (0.140) | (0.131) | (0.142) | (0.141) | (0.195) | (0.188) |
| Bot interaction*After | 0.651*** | 0.455** | 0.339* | 0.393* | 0.076 | 0.159 |
| | (0.165) | (0.156) | (0.168) | (0.167) | (0.224) | (0.214) |
| Sentiment of interaction | | 0.237** | | 0.262** | | 0.319*** |
| | | (0.081) | | (0.089) | | (0.096) |
| Number of retweet of interaction | | −0.0002 | | −0.0002 | | 0.001 |
| | | (0.0004) | | (0.0004) | | (0.0005) |
| Number of likes of interaction | | 0.133*** | | 0.167*** | | 0.097** |
| | | (0.012) | | (0.024) | | (0.032) |
| Topic 1 (Football game protest) | | 0.185 | | 0.041 | | −0.213 |
| | | (0.195) | | (0.238) | | (0.350) |
| Topic 2 (COP25 protest) | | 1.087*** | | 0.947*** | | 1.023** |
| | | (0.209) | | (0.245) | | (0.352) |
| Topic 3 (Anti-XR messages) | | −0.590** | | −1.172*** | | −2.618*** |
| | | (0.212) | | (0.251) | | (0.373) |
| Topic 4 (XR founder's remark on holocaust) | | 0.414* | | 0.122 | | −1.280*** |
| | | (0.202) | | (0.236) | | (0.354) |
| Topic 5 (Disruptive engagement) | | 0.399* | | 0.355 | | −1.206*** |
| | | (0.194) | | (0.234) | | (0.340) |
| Topic 6 (Anti London XR protest messages) | | 0.480* | | 0.416 | | 0.272 |
| | | (0.211) | | (0.249) | | (0.346) |
| Topic 7 (Politicized activism ) | | 0.212 | | 0.114 | | 0.164 |
| | | (0.206) | | (0.241) | | (0.340) |
| Burstiness | | −0.657*** | | −0.644*** | | −0.208*** |
| | | (0.047) | | (0.049) | | (0.056) |
| Observations | 7,051 | 6,660 | 6,391 | 6,000 | 3,931 | 3,660 |
| Log Likelihood | −7,524.297 | −6,852.443 | −6,532.538 | −5,988.802 | −3,761.728 | −3,346.734 |
| Akaike Inf. Crit. | 15,056.590 | 13,734.890 | 13,073.080 | 12,007.600 | 7,531.455 | 6,723.467 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.27:** Linear DiD Model Predicting the Average Sentiment of Tweets 30 days after interaction with astroturfer bots

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| Constant | −0.170*** | −0.171*** | −0.178*** | −0.121*** | −0.182*** | 0.188** |
| | (0.014) | (0.028) | (0.015) | (0.034) | (0.022) | (0.058) |
| Bot interaction (yes = 1) | 0.133*** | 0.097*** | 0.139*** | 0.107*** | 0.126*** | 0.017 |
| | (0.016) | (0.016) | (0.017) | (0.017) | (0.024) | (0.025) |
| After (yes = 1) | 0.109*** | 0.105*** | 0.120*** | 0.115*** | 0.147*** | 0.139*** |
| | (0.019) | (0.019) | (0.021) | (0.021) | (0.031) | (0.030) |
| Bot interaction*After | −0.113*** | −0.107*** | −0.116*** | −0.109*** | −0.140*** | −0.122*** |
| | (0.023) | (0.023) | (0.024) | (0.024) | (0.034) | (0.034) |
| Sentiment of interaction | | −0.0004 | | 0.035** | | 0.0003 |
| | | (0.012) | | (0.013) | | (0.015) |
| Number of retweet of interaction | | 0.0001 | | 0.0001 | | −0.0003*** |
| | | (0.0001) | | (0.0001) | | (0.0001) |
| Number of likes of interaction | | −0.004* | | −0.021*** | | 0.005 |
| | | (0.002) | | (0.004) | | (0.005) |
| Topic 1 (Football game protest) | | 0.030 | | −0.013 | | −0.382*** |
| | | (0.028) | | (0.034) | | (0.057) |
| Topic 2 (COP25 protest) | | −0.065* | | −0.089* | | −0.403*** |
| | | (0.030) | | (0.035) | | (0.058) |
| Topic 3 (Anti-XR messages) | | 0.001 | | −0.090** | | −0.478*** |
| | | (0.030) | | (0.034) | | (0.055) |
| Topic 4 (XR founder's remark on holocaust) | | 0.100*** | | 0.086** | | −0.156** |
| | | (0.029) | | (0.033) | | (0.056) |
| Topic 5 (Disruptive engagement) | | −0.061* | | −0.116*** | | −0.396*** |
| | | (0.028) | | (0.033) | | (0.054) |
| Topic 6 (Anti London XR protest messages) | | 0.150*** | | 0.098** | | −0.084 |
| | | (0.031) | | (0.035) | | (0.056) |
| Topic 7 (Politicized activism ) | | 0.109*** | | 0.061 | | −0.257*** |
| | | (0.030) | | (0.034) | | (0.055) |
| Burstiness | | 0.051*** | | 0.047*** | | 0.036*** |
| | | (0.006) | | (0.006) | | (0.008) |
| Observations | 7,051 | 6,660 | 6,391 | 6,000 | 3,931 | 3,660 |
| $R^2$ | 0.011 | 0.047 | 0.012 | 0.055 | 0.009 | 0.103 |
| Adjusted $R^2$ | 0.010 | 0.045 | 0.012 | 0.053 | 0.008 | 0.100 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.28:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction for users who Support, Neutral or Against XR, botometer CAP=.65

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | Pro | Pro | Neutral | Neutral | Con | Con |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.248*** | −0.822** | −2.338*** | −3.156*** | −1.373*** | 0.330 |
| | (0.056) | (0.258) | (0.196) | (0.578) | (0.154) | (0.256) |
| Bot interaction (yes = 1) | −0.561*** | −0.675*** | 0.995*** | 0.693** | 0.985*** | 1.907*** |
| | (0.077) | (0.084) | (0.230) | (0.231) | (0.171) | (0.187) |
| After (yes = 1) | −0.086 | −0.012 | 0.182 | 0.033 | −0.342 | −0.452* |
| | (0.079) | (0.080) | (0.271) | (0.266) | (0.226) | (0.214) |
| Bot interaction*after | 0.494*** | 0.255* | 0.397 | 0.263 | −0.031 | 0.168 |
| | (0.108) | (0.111) | (0.318) | (0.310) | (0.249) | (0.234) |
| Sentiment of interaction | | −0.412*** | | −0.794*** | | 0.352** |
| | | (0.057) | | (0.172) | | (0.125) |
| Opinion of user interaction | | 0.242*** | | 0.439* | | 0.831*** |
| | | (0.060) | | (0.195) | | (0.125) |
| Number of retweet of interaction | | 0.003*** | | 0.0004 | | −0.0003 |
| | | (0.0003) | | (0.001) | | (0.0003) |
| Number of likes of interaction | | −0.015*** | | 0.089** | | 0.0005 |
| | | (0.005) | | (0.030) | | (0.015) |
| Bot score (astroturf) | | 0.279*** | | 0.049 | | −0.176*** |
| | | (0.027) | | (0.068) | | (0.044) |
| Topic 1 (Football game protest) | | 0.018 | | 0.606 | | −0.910*** |
| | | (0.257) | | (0.546) | | (0.233) |
| Topic 2 (COP25 protest) | | 0.854*** | | 0.662 | | −2.258*** |
| | | (0.257) | | (0.514) | | (0.255) |
| Topic 3 (Anti-XR messages) | | −0.285 | | 1.560** | | −3.164*** |
| | | (0.267) | | (0.504) | | (0.387) |
| Topic 4 (XR founder's remark on holocaust) | | 0.930*** | | −0.440 | | −1.899*** |
| | | (0.257) | | (0.541) | | (0.256) |
| Topic 5 (Disruptive engagement) | | 0.584* | | 0.927 | | −2.792*** |
| | | (0.252) | | (0.534) | | (0.264) |
| Topic 6 (Anti London XR protest messages) | | 0.245 | | −0.926 | | −2.504*** |
| | | (0.253) | | (0.613) | | (0.279) |
| Topic 7 (Politicized activism ) | | −0.435 | | 0.708 | | −3.410*** |
| | | (0.259) | | (0.556) | | (0.296) |
| Burstiness | | 0.060 | | 0.177* | | −0.059 |
| | | (0.031) | | (0.076) | | (0.078) |
| Observations | 10,398 | 9,420 | 3,240 | 2,880 | 3,960 | 3,660 |
| Log Likelihood | −11,905.070 | −10,472.790 | −1,654.172 | −1,520.042 | −3,138.700 | −2,779.994 |
| Akaike Inf. Crit. | 23,818.150 | 20,979.580 | 3,316.344 | 3,074.083 | 6,285.401 | 5,593.989 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.29:** Linear DiD Model Predicting the Average Sentiment of Tweets 30 days after interaction for users who Support, Neutral or Against XR, botometer CAP=.65

| | Sentiment | | | | | |
|---|---|---|---|---|---|---|
| | Pro | Pro | Neutral | Neutral | Con | Con |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.142*** | −0.032 | −0.041* | −0.032 | −0.108*** | 0.064 |
| | (0.009) | (0.065) | (0.019) | (0.065) | (0.021) | (0.041) |
| Bot interaction (yes = 1) | 0.148*** | −0.007 | −0.003 | −0.007 | 0.050* | 0.051* |
| | (0.012) | (0.024) | (0.023) | (0.024) | (0.023) | (0.024) |
| After (yes = 1) | 0.082*** | 0.287*** | 0.267*** | 0.287*** | −0.054 | −0.054* |
| | (0.013) | (0.027) | (0.026) | (0.027) | (0.029) | (0.026) |
| Bot interaction*after | −0.110*** | −0.248*** | −0.236*** | −0.248*** | 0.027 | 0.036 |
| | (0.017) | (0.033) | (0.033) | (0.033) | (0.033) | (0.030) |
| Sentiment of interaction | | −0.102*** | | −0.102*** | | 0.285*** |
| | | (0.020) | | (0.020) | | (0.018) |
| Opinion of user interaction | | 0.157*** | | 0.157*** | | −0.042* |
| | | (0.019) | | (0.019) | | (0.020) |
| Number of retweet of interaction | | −0.0004*** | | −0.0004*** | | −0.0001 |
| | | (0.0001) | | (0.0001) | | (0.0001) |
| Number of likes of interaction | | −0.012*** | | −0.012*** | | −0.014*** |
| | | (0.003) | | (0.003) | | (0.002) |
| Bot score (astroturf) | | −0.053*** | | −0.053*** | | −0.020** |
| | | (0.008) | | (0.008) | | (0.007) |
| Topic 1 (Football game protest) | | 0.037 | | 0.037 | | −0.025 |
| | | (0.062) | | (0.062) | | (0.037) |
| Topic 2 (COP25 protest) | | 0.137* | | 0.137* | | −0.097* |
| | | (0.058) | | (0.058) | | (0.038) |
| Topic 3 (Anti-XR messages) | | 0.135* | | 0.135* | | −0.064 |
| | | (0.057) | | (0.057) | | (0.048) |
| Topic 4 (XR founder's remark on holocaust) | | 0.142* | | 0.142* | | −0.157*** |
| | | (0.059) | | (0.059) | | (0.038) |
| Topic 5 (Disruptive engagement) | | 0.041 | | 0.041 | | −0.255*** |
| | | (0.060) | | (0.060) | | (0.037) |
| Topic 6 (Anti London XR protest messages) | | 0.229*** | | 0.229*** | | −0.286*** |
| | | (0.062) | | (0.062) | | (0.041) |
| Topic 7 (Politicized activism ) | | 0.271*** | | 0.271*** | | 0.013 |
| | | (0.062) | | (0.062) | | (0.041) |
| Burstiness | | 0.013 | | 0.013 | | 0.012 |
| | | (0.009) | | (0.009) | | (0.011) |
| Observations | 10,398 | 2,880 | 3,240 | 2,880 | 3,960 | 3,660 |
| $R^2$ | 0.016 | 0.142 | 0.048 | 0.142 | 0.005 | 0.115 |
| Adjusted $R^2$ | 0.015 | 0.137 | 0.047 | 0.137 | 0.005 | 0.111 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.30:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction for users who Support, Neutral or Against XR, botometer CAP=.70

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | Pro | Pro | Neutral | Neutral | Con | Con |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.219*** | −3.511*** | −2.303*** | −3.511*** | −1.455*** | −0.203 |
| | (0.060) | (0.531) | (0.197) | (0.531) | (0.170) | (0.264) |
| Bot interaction (yes = 1) | −0.644*** | 0.944*** | 0.588* | 0.944*** | 0.856*** | 1.326*** |
| | (0.083) | (0.234) | (0.231) | (0.234) | (0.185) | (0.206) |
| After (yes = 1) | −0.065 | 0.120 | 0.236 | 0.120 | −0.241 | −0.407 |
| | (0.085) | (0.259) | (0.271) | (0.259) | (0.247) | (0.233) |
| Bot interaction*After | 0.190 | −0.290 | −0.274 | −0.290 | −0.067 | 0.156 |
| | (0.118) | (0.309) | (0.320) | (0.309) | (0.269) | (0.252) |
| Sentiment of interaction | | −1.324*** | | −1.324*** | | 0.031 |
| | | (0.205) | | (0.205) | | (0.130) |
| Opinion of user interaction | | 0.121 | | 0.121 | | 0.323* |
| | | (0.230) | | (0.230) | | (0.126) |
| Number of retweet of interaction | | 0.0004 | | 0.0004 | | −0.0001 |
| | | (0.001) | | (0.001) | | (0.0004) |
| Number of likes of interaction | | 0.0001 | | 0.0001 | | 0.021 |
| | | (0.034) | | (0.034) | | (0.015) |
| Bot score (astroturf) | | 0.200** | | 0.200** | | 0.014 |
| | | (0.075) | | (0.075) | | (0.045) |
| Topic 1 (Football game protest) | | 0.651 | | 0.651 | | −0.585* |
| | | (0.510) | | (0.510) | | (0.259) |
| Topic 2 (COP25 protest) | | 0.416 | | 0.416 | | −1.896*** |
| | | (0.447) | | (0.447) | | (0.286) |
| Topic 3 (Anti-XR messages) | | 0.584 | | 0.584 | | −2.312*** |
| | | (0.455) | | (0.455) | | (0.393) |
| Topic 4 (XR founder's remark on holocaust) | | −0.503 | | −0.503 | | −2.762*** |
| | | (0.483) | | (0.483) | | (0.294) |
| Topic 5 (Disruptive engagement) | | 1.274** | | 1.274** | | −2.482*** |
| | | (0.473) | | (0.473) | | (0.285) |
| Topic 6 (Anti London XR protest messages) | | −0.747 | | −0.747 | | −2.283*** |
| | | (0.592) | | (0.592) | | (0.305) |
| Topic 7 (Politicized activism ) | | 0.857 | | 0.857 | | −2.820*** |
| | | (0.506) | | (0.506) | | (0.312) |
| Burstiness | | 0.019 | | 0.019 | | −0.087 |
| | | (0.085) | | (0.085) | | (0.082) |
| Observations | 8,778 | 2,460 | 2,700 | 2,460 | 3,600 | 3,300 |
| Log Likelihood | −9,543.139 | −1,040.597 | −1,134.655 | −1,040.597 | −2,713.683 | −2,364.884 |
| Akaike Inf. Crit. | 19,094.280 | 2,115.195 | 2,277.310 | 2,115.195 | 5,435.366 | 4,763.769 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.31:** Linear DiD Model Predicting the Average Sentiment of Tweets 30 days after interaction for users who Support, Neutral or Against XR, botometer CAP=.70

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | Pro | Pro | Neutral | Neutral | Con | Con |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.147*** | 0.028 | −0.063** | 0.028 | −0.134*** | 0.063 |
| | (0.010) | (0.068) | (0.021) | (0.068) | (0.023) | (0.046) |
| Bot interaction (yes = 1) | 0.141*** | 0.010 | 0.004 | 0.010 | 0.061* | 0.044 |
| | (0.013) | (0.027) | (0.026) | (0.027) | (0.026) | (0.028) |
| After (yes = 1) | 0.075*** | 0.328*** | 0.306*** | 0.328*** | −0.031 | −0.031 |
| | (0.014) | (0.029) | (0.030) | (0.029) | (0.033) | (0.030) |
| Bot interaction*After | −0.077*** | −0.279*** | −0.267*** | −0.279*** | 0.009 | 0.020 |
| | (0.018) | (0.036) | (0.036) | (0.036) | (0.037) | (0.034) |
| Sentiment of interaction | | −0.081*** | | −0.081*** | | 0.267*** |
| | | (0.024) | | (0.024) | | (0.020) |
| Opinion of user interaction | | 0.155*** | | 0.155*** | | −0.044* |
| | | (0.021) | | (0.021) | | (0.021) |
| Number of retweet of interaction | | −0.0004*** | | −0.0004*** | | −0.0001* |
| | | (0.0001) | | (0.0001) | | (0.0001) |
| Number of likes of interaction | | −0.020*** | | −0.020*** | | −0.011*** |
| | | (0.003) | | (0.003) | | (0.002) |
| Bot score (astroturf) | | −0.073*** | | −0.073*** | | −0.013 |
| | | (0.010) | | (0.010) | | (0.007) |
| Topic 1 (Football game protest) | | 0.151* | | 0.151* | | −0.080 |
| | | (0.066) | | (0.066) | | (0.045) |
| Topic 2 (COP25 protest) | | 0.133* | | 0.133* | | −0.133** |
| | | (0.060) | | (0.060) | | (0.047) |
| Topic 3 (Anti-XR messages) | | 0.077 | | 0.077 | | −0.172** |
| | | (0.060) | | (0.060) | | (0.056) |
| Topic 4 (XR founder's remark on holocaust) | | 0.137* | | 0.137* | | −0.197*** |
| | | (0.060) | | (0.060) | | (0.045) |
| Topic 5 (Disruptive engagement) | | 0.018 | | 0.018 | | −0.271*** |
| | | (0.062) | | (0.062) | | (0.045) |
| Topic 6 (Anti London XR protest messages) | | 0.210** | | 0.210** | | −0.307*** |
| | | (0.064) | | (0.064) | | (0.049) |
| Topic 7 (Politicized activism ) | | 0.210** | | 0.210** | | −0.009 |
| | | (0.066) | | (0.066) | | (0.048) |
| Burstiness | | 0.007 | | 0.007 | | 0.038** |
| | | (0.010) | | (0.010) | | (0.013) |
| Observations | 8,778 | 2,460 | 2,700 | 2,460 | 3,600 | 3,300 |
| $R^2$ | 0.017 | 0.169 | 0.057 | 0.169 | 0.004 | 0.098 |
| Adjusted $R^2$ | 0.017 | 0.164 | 0.056 | 0.164 | 0.004 | 0.094 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table A.32:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction for users who Support, Neutral or Against XR, botometer CAP=.75

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | | | Amount | | | |
| | Pro | Pro | Neutral | Neutral | Con | Con |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.060 | −3.378*** | −2.568*** | −3.378*** | −1.415*** | 0.334 |
| | (0.069) | (0.482) | (0.282) | (0.482) | (0.206) | (0.297) |
| Bot interaction (yes = 1) | −0.487*** | 0.716* | 0.758* | 0.716* | 0.306 | 1.256*** |
| | (0.097) | (0.319) | (0.324) | (0.319) | (0.223) | (0.279) |
| After (yes = 1) | 0.082 | 0.084 | 0.197 | 0.084 | −0.061 | −0.153 |
| | (0.097) | (0.365) | (0.388) | (0.365) | (0.294) | (0.292) |
| Bot interaction*After | −0.097 | −0.486 | −0.294 | −0.486 | −0.055 | −0.020 |
| | (0.138) | (0.427) | (0.451) | (0.427) | (0.317) | (0.311) |
| Sentiment of interaction | | −1.588*** | | −1.588*** | | 0.710*** |
| | | (0.297) | | (0.297) | | (0.171) |
| Opinion of user interaction | | 0.518 | | 0.518 | | 0.017 |
| | | (0.266) | | (0.266) | | (0.183) |
| Number of retweet of interaction | | −0.004 | | −0.004 | | −0.001* |
| | | (0.002) | | (0.002) | | (0.0004) |
| Number of likes of interaction | | −0.013 | | −0.013 | | 0.003 |
| | | (0.039) | | (0.039) | | (0.017) |
| Bot score (astroturf) | | 0.252* | | 0.252* | | 0.104 |
| | | (0.102) | | (0.102) | | (0.054) |
| Topic 1 (Football game protest) | | | | | | −1.866*** |
| | | | | | | (0.325) |
| Topic 2 (COP25 protest) | | 0.354 | | 0.354 | | −2.300*** |
| | | (0.446) | | (0.446) | | (0.345) |
| Topic 3 (Anti-XR messages) | | 0.886* | | 0.886* | | −3.175*** |
| | | (0.421) | | (0.421) | | (0.411) |
| Topic 4 (XR founder's remark on holocaust) | | −0.637 | | −0.637 | | −3.689*** |
| | | (0.492) | | (0.492) | | (0.363) |
| Topic 5 (Disruptive engagement) | | 1.694*** | | 1.694*** | | −3.877*** |
| | | (0.451) | | (0.451) | | (0.365) |
| Topic 6 (Anti London XR protest messages) | | −0.329 | | −0.329 | | −3.233*** |
| | | (0.519) | | (0.519) | | (0.386) |
| Topic 7 (Politicized activism ) | | 0.668 | | 0.668 | | −3.562*** |
| | | (0.521) | | (0.521) | | (0.368) |
| Burstiness | | 0.167 | | 0.167 | | −0.014 |
| | | (0.116) | | (0.116) | | (0.104) |
| Observations | 6,198 | 1,680 | 1,920 | 1,680 | 2,640 | 2,340 |
| Log Likelihood | −6,622.960 | −617.740 | −699.774 | −617.740 | −1,740.707 | −1,450.346 |
| Akaike Inf. Crit. | 13,253.920 | 1,267.480 | 1,407.548 | 1,267.480 | 3,489.414 | 2,934.692 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.33:** Linear DiD Model Predicting the Average Sentiment of Tweets 30 days after interaction for users who Support, Neutral or Against XR, botometer CAP=.75

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | Pro | Pro | Neutral | Neutral | Con | Con |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.133*** | −0.204*** | 0.010 | −0.204*** | −0.162*** | 0.234*** |
| | (0.011) | (0.052) | (0.026) | (0.052) | (0.031) | (0.061) |
| Bot interaction (yes = 1) | 0.098*** | −0.002 | −0.073* | −0.002 | 0.081* | 0.063 |
| | (0.015) | (0.030) | (0.031) | (0.030) | (0.034) | (0.036) |
| After (yes = 1) | 0.076*** | 0.356*** | 0.320*** | 0.356*** | 0.099* | 0.099* |
| | (0.016) | (0.033) | (0.036) | (0.033) | (0.044) | (0.039) |
| Bot interaction*After | −0.040 | −0.320*** | −0.297*** | −0.320*** | −0.114* | −0.099* |
| | (0.022) | (0.041) | (0.044) | (0.041) | (0.048) | (0.043) |
| Sentiment of interaction | | −0.388*** | | −0.388*** | | 0.273*** |
| | | (0.031) | | (0.031) | | (0.026) |
| Opinion of user interaction | | 0.276*** | | 0.276*** | | −0.130*** |
| | | (0.022) | | (0.022) | | (0.030) |
| Number of retweet of interaction | | 0.001** | | 0.001** | | −0.0003*** |
| | | (0.0002) | | (0.0002) | | (0.0001) |
| Number of likes of interaction | | −0.033*** | | −0.033*** | | −0.012*** |
| | | (0.004) | | (0.004) | | (0.003) |
| Bot score (astroturf) | | −0.016 | | −0.016 | | −0.031*** |
| | | (0.011) | | (0.011) | | (0.008) |
| Topic 1 (Football game protest) | | | | | | −0.291*** |
| | | | | | | (0.063) |
| Topic 2 (COP25 protest) | | 0.013 | | 0.013 | | −0.367*** |
| | | (0.050) | | (0.050) | | (0.065) |
| Topic 3 (Anti-XR messages) | | 0.351*** | | 0.351*** | | −0.412*** |
| | | (0.048) | | (0.048) | | (0.069) |
| Topic 4 (XR founder's remark on holocaust) | | 0.152** | | 0.152** | | −0.355*** |
| | | (0.047) | | (0.047) | | (0.063) |
| Topic 5 (Disruptive engagement) | | 0.315*** | | 0.315*** | | −0.456*** |
| | | (0.051) | | (0.051) | | (0.065) |
| Topic 6 (Anti London XR protest messages) | | 0.479*** | | 0.479*** | | −0.527*** |
| | | (0.049) | | (0.049) | | (0.068) |
| Topic 7 (Politicized activism ) | | 0.546*** | | 0.546*** | | −0.159* |
| | | (0.054) | | (0.054) | | (0.064) |
| Burstiness | | −0.018 | | −0.018 | | 0.085*** |
| | | (0.011) | | (0.011) | | (0.017) |
| Observations | 6,198 | 1,680 | 1,920 | 1,680 | 2,640 | 2,340 |
| $R^2$ | 0.013 | 0.307 | 0.087 | 0.307 | 0.002 | 0.122 |
| Adjusted $R^2$ | 0.012 | 0.300 | 0.086 | 0.300 | 0.001 | 0.116 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.34:** Linear DiD Model Predicting the Change in Support 30 days after interaction for users who Support, Neutral or Against XR, with varied botometer CAP scores, without controls

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Opinion Change | | | | | | | | |
| | Pro | Neutral | Con | Pro | Neutral | Con | Pro | Neutral | Con |
| Constant | $0.172^{***}$ | $0.105^{***}$ | $-0.520^{***}$ | $0.204^{***}$ | $0.133^{***}$ | $-0.583^{***}$ | $0.204^{***}$ | $0.133^{***}$ | $-0.857^{***}$ |
| | (0.009) | (0.014) | (0.026) | (0.010) | (0.016) | (0.029) | (0.010) | (0.016) | (0.037) |
| Bot interaction (yes = 1) | −0.008 | $-0.220^{***}$ | $0.195^{***}$ | $-0.051^{***}$ | $-0.217^{***}$ | $0.265^{***}$ | $-0.051^{***}$ | $-0.217^{***}$ | $0.525^{***}$ |
| | (0.012) | (0.018) | (0.030) | (0.013) | (0.020) | (0.032) | (0.013) | (0.020) | (0.041) |
| After (yes = 1) | −0.001 | −0.000 | −0.000 | −0.001 | 0.000 | −0.000 | −0.001 | 0.000 | −0.000 |
| | (0.013) | (0.020) | (0.037) | (0.014) | (0.023) | (0.040) | (0.014) | (0.023) | (0.053) |
| Bot interaction*After | −0.003 | 0.000 | 0.000 | −0.004 | −0.000 | 0.000 | −0.004 | −0.000 | 0.000 |
| | (0.017) | (0.025) | (0.042) | (0.019) | (0.028) | (0.045) | (0.019) | (0.028) | (0.058) |
| CAP | 65 | 70 | 75 | 65 | 70 | 75 | 65 | 70 | 75 |
| Observations | 10,398 | 3,240 | 3,960 | 8,778 | 2,700 | 3,600 | 8,778 | 2,700 | 2,640 |
| $R^2$ | 0.0001 | 0.086 | 0.021 | 0.004 | 0.082 | 0.037 | 0.004 | 0.082 | 0.112 |
| Adjusted $R^2$ | −0.0002 | 0.085 | 0.021 | 0.003 | 0.081 | 0.036 | 0.003 | 0.081 | 0.111 |
| *Note:* | | | | | | | | | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 |

**Table A.35:** Linear DiD Model Predicting the Change in Support 30 days after interaction for users who Support, Neutral or Against XR, with varied botometer CAP scores, with controls

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Opinion Change | | | | | | | | |
| | Pro | Neutral | Con | Pro | Neutral | Con | Pro | Neutral | Con |
| Constant | 0.819*** | −0.793*** | −0.133** | 0.868*** | −0.798*** | −0.061 | 0.868*** | −0.798*** | −0.031 |
| | (0.040) | (0.048) | (0.051) | (0.041) | (0.042) | (0.056) | (0.041) | (0.042) | (0.068) |
| Bot interaction (yes = 1) | 0.070*** | −0.214*** | 0.389*** | 0.026 | −0.106*** | 0.490*** | 0.026 | −0.106*** | 0.705*** |
| | (0.013) | (0.018) | (0.030) | (0.014) | (0.016) | (0.034) | (0.014) | (0.016) | (0.039) |
| After (yes = 1) | −0.000 | 0.000 | −0.000 | 0.000 | −0.000 | 0.000 | 0.000 | −0.000 | −0.000 |
| | (0.013) | (0.020) | (0.033) | (0.014) | (0.018) | (0.036) | (0.014) | (0.018) | (0.043) |
| Bot interaction*After | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | −0.000 | 0.000 | 0.000 | 0.000 |
| | (0.017) | (0.024) | (0.038) | (0.019) | (0.022) | (0.041) | (0.019) | (0.022) | (0.048) |
| Sentiment of interaction | −0.044*** | −0.081*** | −0.328*** | −0.040*** | −0.136*** | −0.214*** | −0.040*** | −0.136*** | −0.067* |
| | (0.009) | (0.015) | (0.022) | (0.010) | (0.015) | (0.024) | (0.010) | (0.015) | (0.028) |
| Opinion of user interaction | −0.027** | −0.073*** | 0.243*** | −0.027** | 0.022 | 0.243*** | −0.027** | 0.022 | −0.250*** |
| | (0.009) | (0.014) | (0.025) | (0.010) | (0.013) | (0.026) | (0.010) | (0.013) | (0.033) |
| Num. of retweet of interaction | −0.0003*** | 0.00003 | −0.001*** | −0.0004*** | 0.0004*** | −0.001*** | −0.0004*** | 0.0004*** | −0.001*** |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Num. of likes of interaction | 0.008*** | 0.020*** | 0.037*** | 0.007*** | 0.026*** | 0.030*** | 0.007*** | 0.026*** | 0.019*** |
| | (0.001) | (0.002) | (0.003) | (0.001) | (0.002) | (0.003) | (0.001) | (0.002) | (0.003) |
| Bot score (astroturf) | −0.090*** | −0.003 | 0.012 | −0.099*** | −0.052*** | −0.023** | −0.099*** | −0.052*** | −0.027** |
| | (0.004) | (0.006) | (0.008) | (0.005) | (0.006) | (0.009) | (0.005) | (0.006) | (0.009) |
| Topic 1 (Football game protest) | −0.614*** | 1.078*** | −0.385*** | −0.537*** | 1.396*** | −0.417*** | −0.537*** | 1.396*** | −0.662*** |
| | (0.040) | (0.046) | (0.046) | (0.042) | (0.041) | (0.054) | (0.042) | (0.041) | (0.070) |
| Topic 2 (COP25 protest) | −0.518*** | 0.723*** | −0.503*** | −0.540*** | 0.836*** | −0.572*** | −0.540*** | 0.836*** | −1.076*** |
| | (0.040) | (0.043) | (0.047) | (0.041) | (0.036) | (0.057) | (0.041) | (0.036) | (0.072) |
| Topic 3 (Anti-XR messages) | −0.292*** | 0.868*** | −0.602*** | −0.188*** | 0.872*** | −0.748*** | −0.188*** | 0.872*** | −0.921*** |
| | (0.041) | (0.043) | (0.059) | (0.042) | (0.037) | (0.067) | (0.042) | (0.037) | (0.076) |
| Topic 4 (XR founder's remark on holocaust) | −0.562*** | 0.928*** | −0.646*** | −0.552*** | 0.852*** | −0.592*** | −0.552*** | 0.852*** | −0.988*** |
| | (0.040) | (0.044) | (0.047) | (0.041) | (0.037) | (0.055) | (0.041) | (0.037) | (0.069) |
| Topic 5 (Disruptive engagement) | −0.415*** | 0.887*** | −0.665*** | −0.418*** | 0.934*** | −0.745*** | −0.418*** | 0.934*** | −1.368*** |
| | (0.039) | (0.045) | (0.046) | (0.040) | (0.038) | (0.055) | (0.040) | (0.038) | (0.071) |
| Topic 6 (Anti London XR protest messages) | −0.540*** | 0.803*** | −0.504*** | −0.539*** | 0.860*** | −0.583*** | −0.539*** | 0.860*** | −1.180*** |
| | (0.039) | (0.046) | (0.050) | (0.040) | (0.039) | (0.059) | (0.040) | (0.039) | (0.075) |
| Topic 7 (Politicized activism ) | −0.451*** | 0.978*** | −0.621*** | −0.464*** | 0.986*** | −0.675*** | −0.464*** | 0.986*** | −0.787*** |
| | (0.040) | (0.046) | (0.050) | (0.040) | (0.040) | (0.058) | (0.040) | (0.040) | (0.070) |
| Burstiness | 0.005 | 0.025*** | −0.014 | 0.009 | 0.094*** | −0.135*** | 0.009 | 0.094*** | −0.213*** |
| | (0.005) | (0.006) | (0.013) | (0.005) | (0.006) | (0.015) | (0.005) | (0.006) | (0.019) |
| CAP | 65 | 70 | 75 | 65 | 70 | 75 | 65 | 70 | 75 |
| Observations | 9,420 | 2,880 | 3,660 | 7,920 | 2,460 | 3,300 | 7,920 | 2,460 | 2,340 |
| $R^2$ | 0.118 | 0.348 | 0.238 | 0.132 | 0.537 | 0.229 | 0.132 | 0.537 | 0.420 |
| Adjusted $R^2$ | 0.117 | 0.344 | 0.234 | 0.130 | 0.534 | 0.225 | 0.130 | 0.534 | 0.416 |
| *Note:* | | | | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table A.36:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction with astroturfing bots, Coarsened Exact Matching only

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Bot interaction | −0.064 | −0.346** | −0.380** | −0.465*** | −0.408* | −0.530** |
| | (0.134) | (0.123) | (0.125) | (0.127) | (0.171) | (0.174) |
| After | −0.081 | −0.156 | −0.058 | −0.163 | −0.087 | −0.389 |
| | (0.166) | (0.146) | (0.154) | (0.152) | (0.212) | (0.212) |
| Sentiment of Interaction | | 0.480*** | | 0.748*** | | 0.677*** |
| | | (0.083) | | (0.088) | | (0.103) |
| Number of retweets | | −0.004*** | | −0.008*** | | −0.006** |
| | | (0.001) | | (0.001) | | (0.002) |
| Number of likes | | 0.212*** | | 0.078** | | 0.181*** |
| | | (0.016) | | (0.026) | | (0.032) |
| Topic 1 (Football game protest) | | 0.658** | | 0.351 | | 0.147 |
| | | (0.202) | | (0.216) | | (0.343) |
| Topic 2 (COP25 protest) | | 1.017*** | | 0.660** | | −1.044* |
| | | (0.230) | | (0.238) | | (0.430) |
| Topic 3 (Anti-XR messages) | | −0.011 | | −0.688** | | −1.990*** |
| | | (0.221) | | (0.255) | | (0.417) |
| Topic 4 (XR founder's remark on holocaust) | | −0.311 | | −0.423 | | −0.894* |
| | | (0.212) | | (0.226) | | (0.349) |
| Topic 5 (Disruptive engagement) | | 0.186 | | −0.267 | | −0.929** |
| | | (0.208) | | (0.219) | | (0.346) |
| Topic 6 (Anti London XR protest messages) | | 1.289*** | | 1.004*** | | 0.908** |
| | | (0.213) | | (0.223) | | (0.345) |
| Topic 7 (Politicized activism ) | | 0.313 | | −0.029 | | −0.108 |
| | | (0.214) | | (0.219) | | (0.342) |
| Burstiness | | −0.208*** | | −0.132* | | −0.059 |
| | | (0.049) | | (0.052) | | (0.063) |
| Bot interaction:After | 0.486* | 0.157 | −0.090 | 0.003 | −0.014 | 0.174 |
| | (0.190) | (0.170) | (0.178) | (0.176) | (0.243) | (0.240) |
| Constant | −0.438*** | −1.200*** | −0.374*** | −0.560* | −0.353* | −0.252 |
| | (0.117) | (0.202) | (0.109) | (0.220) | (0.149) | (0.360) |
| Observations | 5,551 | 5,160 | 4,951 | 4,560 | 3,271 | 3,000 |
| Log Likelihood | −5,232.702 | −4,529.911 | −4,206.804 | −3,689.154 | −2,694.084 | −2,331.438 |
| Akaike Inf. Crit. | 10,473.410 | 9,089.822 | 8,421.608 | 7,408.308 | 5,396.168 | 4,692.875 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.37:** Linear DiD Model Predicting the Average Sentiment 30 days after interaction with astroturfing bots, Coarsened Exact Matching only

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Bot interaction | 0.135*** | 0.097*** | 0.127*** | 0.086*** | 0.132*** | 0.037 |
| | (0.019) | (0.020) | (0.021) | (0.022) | (0.027) | (0.028) |
| After | 0.120*** | 0.114*** | 0.115*** | 0.107*** | 0.168*** | 0.160*** |
| | (0.024) | (0.024) | (0.026) | (0.026) | (0.033) | (0.033) |
| Sentiment of Interaction | | −0.019 | | 0.021 | | −0.023 |
| | | (0.013) | | (0.014) | | (0.016) |
| Number of retweets | | 0.0004** | | 0.00004 | | −0.0001 |
| | | (0.0001) | | (0.0002) | | (0.0002) |
| Number of likes | | −0.010*** | | −0.030*** | | −0.001 |
| | | (0.003) | | (0.004) | | (0.005) |
| Topic 1 (Football game protest) | | −0.016 | | −0.034 | | −0.399*** |
| | | (0.032) | | (0.036) | | (0.059) |
| Topic 2 (COP25 protest) | | −0.100** | | −0.078* | | −0.435*** |
| | | (0.036) | | (0.040) | | (0.067) |
| Topic 3 (Anti-XR messages) | | −0.017 | | −0.086* | | −0.485*** |
| | | (0.034) | | (0.040) | | (0.063) |
| Topic 4 (XR founder's remark on holocaust) | | 0.085** | | 0.100** | | −0.170** |
| | | (0.032) | | (0.036) | | (0.059) |
| Topic 5 (Disruptive engagement) | | −0.094** | | −0.111** | | −0.416*** |
| | | (0.032) | | (0.036) | | (0.058) |
| Topic 6 (Anti London XR protest messages) | | 0.075* | | 0.041 | | −0.148* |
| | | (0.034) | | (0.038) | | (0.060) |
| Topic 7 (Politicized activism ) | | 0.078* | | 0.059 | | −0.285*** |
| | | (0.033) | | (0.036) | | (0.058) |
| Burstiness | | 0.044*** | | 0.041*** | | 0.032*** |
| | | (0.007) | | (0.008) | | (0.009) |
| Bot interaction:After | −0.130*** | −0.122*** | −0.115*** | −0.105*** | −0.184*** | −0.166*** |
| | (0.027) | (0.028) | (0.030) | (0.030) | (0.038) | (0.037) |
| Constant | −0.167*** | −0.135*** | −0.162*** | −0.091* | −0.176*** | 0.204*** |
| | (0.017) | (0.032) | (0.018) | (0.036) | (0.024) | (0.061) |
| Observations | 5,551 | 5,160 | 4,951 | 4,560 | 3,271 | 3,000 |
| $R^2$ | 0.009 | 0.044 | 0.008 | 0.052 | 0.009 | 0.099 |
| Adjusted $R^2$ | 0.009 | 0.041 | 0.008 | 0.049 | 0.008 | 0.095 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.38:** Negative Binomial DiD Model Predicting the Average Daily Tweet Count 30 days after interaction with bots, with user region controlled

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Amount | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.056 | 0.141 | −0.188** | 0.602*** | −0.066 | 1.010*** |
| | p = 0.274 | p = 0.358 | p = 0.004 | p = 0.0005 | p = 0.230 | p = 0.00000 |
| Bot interaction | −0.414*** | −0.303*** | −0.608*** | −0.337*** | −0.567*** | −0.266** |
| | p = 0.000 | p = 0.00001 | p = 0.000 | p = 0.00001 | p = 0.000 | p = 0.002 |
| After | −0.090 | −0.124 | 0.078 | −0.115 | −0.066 | −0.061 |
| | p = 0.215 | p = 0.092 | p = 0.395 | p = 0.138 | p = 0.395 | p = 0.498 |
| Bot interaction * After | 0.333*** | 0.237* | −0.126 | 0.045 | 0.059 | 0.028 |
| | p = 0.0004 | p = 0.012 | p = 0.288 | p = 0.659 | p = 0.553 | p = 0.813 |
| Sentiment of Interaction | | −0.090 | | 0.011 | | 0.075 |
| | | p = 0.066 | | p = 0.841 | | p = 0.205 |
| Number of retweets | | 0.001*** | | 0.0003 | | −0.00000 |
| | | p = 0.00000 | | p = 0.230 | | p = 0.988 |
| Number of likes | | −0.003 | | −0.026*** | | −0.031*** |
| | | p = 0.500 | | p = 0.00000 | | p = 0.000 |
| Topic 1 (Football game protest) | | −0.630*** | | −0.776*** | | −1.283*** |
| | | p = 0.00004 | | p = 0.00001 | | p = 0.000 |
| Topic 2 (COP25 protest) | | −0.051 | | −0.524** | | −0.592** |
| | | p = 0.737 | | p = 0.002 | | p = 0.003 |
| Topic 3 (Anti-XR messages) | | −0.463** | | −1.942*** | | −2.493*** |
| | | p = 0.004 | | p = 0.000 | | p = 0.000 |
| Topic 4 (XR founder's remark on holocaust) | | −0.147 | | −0.628*** | | −1.557*** |
| | | p = 0.346 | | p = 0.0003 | | p = 0.000 |
| Topic 5 (Disruptive engagement) | | −0.504*** | | −0.703*** | | −0.829*** |
| | | p = 0.001 | | p = 0.00002 | | p = 0.00002 |
| Topic 6 (Anti London XR protest messages) | | −0.517*** | | −0.894*** | | −1.181*** |
| | | p = 0.001 | | p = 0.00000 | | p = 0.000 |
| Topic 7 (Politicized activism) | | −1.098*** | | −1.323*** | | −1.288*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.000 |
| Australia and NZ | | 1.645*** | | 1.497*** | | 1.128*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.000 |
| Canada | | −0.868*** | | −0.318 | | 0.447 |
| | | p = 0.00001 | | p = 0.085 | | p = 0.122 |
| Europe other | | −0.838*** | | −1.111*** | | −1.181*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.000 |
| Other | | 1.085*** | | −0.462 | | −0.317 |
| | | p = 0.00000 | | p = 0.129 | | p = 0.300 |
| UK and Ireland | | −0.463*** | | −0.469*** | | −0.689*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.000 |
| USA | | 0.448*** | | 0.344*** | | −0.234** |
| | | p = 0.000 | | p = 0.00001 | | p = 0.006 |
| Burstiness | | −0.021 | | −0.203*** | | −0.192*** |
| | | p = 0.454 | | p = 0.000 | | p = 0.00001 |
| Observations | 17,598 | 14,580 | 10,758 | 12,600 | 15,078 | 9,060 |
| Log Likelihood | −17,045.970 | −13,986.710 | −9,297.998 | −11,323.400 | −13,693.760 | −7,967.407 |
| Akaike Inf. Crit. | 34,099.940 | 28,015.410 | 18,604.000 | 22,688.790 | 27,395.530 | 15,976.810 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table A.39:** Linear DiD Model Predicting the Average Sentiment 30 days after interaction with bots, with demographic controlled

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sentiment | | | | | |
| | 65 | 65 | 70 | 70 | 75 | 75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | −0.121*** | −0.244*** | −0.132*** | −0.249*** | −0.114*** | −0.228*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 |
| Bot interaction (yes = 1) | 0.099*** | 0.090*** | 0.095*** | 0.092*** | 0.059*** | 0.056*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.00001 | p = 0.00003 |
| After | 0.095*** | 0.087*** | 0.098*** | 0.097*** | 0.115*** | 0.119*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 |
| Bot interaction: After | −0.112*** | −0.117*** | −0.099*** | −0.113*** | −0.098*** | −0.116*** |
| | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 | p = 0.000 |
| Sentiment of Interaction | | 0.026*** | | 0.019* | | 0.002 |
| | | p = 0.001 | | p = 0.023 | | p = 0.850 |
| Number of retweets | | 0.0002*** | | 0.0001** | | 0.0002*** |
| | | p = 0.00002 | | p = 0.002 | | p = 0.00000 |
| Number of likes | | −0.002*** | | −0.003*** | | −0.003*** |
| | | p = 0.0005 | | p = 0.0003 | | p = 0.0004 |
| Topic 1 (Football game protest) | | 0.094*** | | 0.027 | | 0.025 |
| | | p = 0.0001 | | p = 0.327 | | p = 0.458 |
| Topic 2 (COP25 protest) | | 0.101*** | | 0.089*** | | 0.008 |
| | | p = 0.00004 | | p = 0.001 | | p = 0.810 |
| Topic 3 (Anti-XR messages) | | 0.148*** | | 0.129*** | | 0.122*** |
| | | p = 0.000 | | p = 0.00001 | | p = 0.0005 |
| Topic 4 (XR founder's remark on holocaust) | | 0.057* | | 0.027 | | −0.004 |
| | | p = 0.023 | | p = 0.322 | | p = 0.902 |
| Topic 5 (Disruptive engagement) | | 0.038 | | 0.041 | | 0.038 |
| | | p = 0.113 | | p = 0.121 | | p = 0.254 |
| Topic 6 (Anti London XR protest messages) | | 0.090*** | | 0.081** | | 0.053 |
| | | p = 0.0002 | | p = 0.002 | | p = 0.103 |
| Topic 7 (Politicized activism ) | | 0.186*** | | 0.188*** | | 0.146*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.00002 |
| Australia and NZ | | −0.022 | | −0.020 | | 0.001 |
| | | p = 0.166 | | p = 0.269 | | p = 0.953 |
| Canada | | −0.015 | | 0.003 | | 0.243*** |
| | | p = 0.585 | | p = 0.912 | | p = 0.00000 |
| Europe other | | 0.133*** | | 0.150*** | | 0.107*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.00001 |
| Other | | −0.049 | | −0.153*** | | −0.123** |
| | | p = 0.150 | | p = 0.0002 | | p = 0.003 |
| UK and Ireland | | 0.084*** | | 0.082*** | | 0.114*** |
| | | p = 0.000 | | p = 0.000 | | p = 0.000 |
| USA | | 0.011 | | 0.023* | | 0.047*** |
| | | p = 0.333 | | p = 0.047 | | p = 0.001 |
| Burstiness | | 0.007 | | 0.004 | | −0.015* |
| | | p = 0.083 | | p = 0.372 | | p = 0.019 |
| Observations | 17,598 | 14,580 | 15,078 | 12,600 | 10,758 | 9,060 |
| $R^2$ | 0.007 | 0.031 | 0.007 | 0.035 | 0.007 | 0.042 |
| Adjusted $R^2$ | 0.007 | 0.029 | 0.007 | 0.034 | 0.006 | 0.040 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

# B

# Supplementary information for Chapter 5

This appendix provides supplementary information for Chapter 5 of this dissertation. It details the methodological procedures, rationale for method selection, robustness checks, and full model tables across the following aspects: data collection, bot identification, support group categorisation, topic modelling and network construction.

## Data collection

For our investigation into human and bot networks related to the Black Lives Matter (BLM) protests on X (formerly Twitter), we concentrated on the period from June 6 to 8 June, 2020. This interval was identified as the peak of protest-related activity following George Floyd' s death on 25 May, 2020.

We made use of the Twitter Academic Product track API to collect tweets from this period, allowing us to obtain a comprehensive archive of tweets containing keywords such as 'BLM,' 'George Floyd,' and several variations with minor spelling differences. The full list of keywords employed is provided in Table B.1.

Along with tweet content, we collected associated metadata, including user IDs. We then used these IDs to fetch additional user information via the Twitter user lookup API, which was available by the time of data collection. Our final dataset

**Table B.1:** Keywords for tweet collection.

| **Keywords** |
|---|
| BlackLivesMatter |
| black lives matter |
| BLM |
| George Floyd |
| georgefloyd |

includes over 2,500,000 tweets from more than 1,000,000 users within the specified time frame, from June 6, 2020, 00:00 to June 8, 2020, 23:59.

Given that the API provided a complete archive of tweets, we consider our sample to be relatively comprehensive for capturing all relevant communication related to the Black Lives Matter (BLM) protests on X (formerly Twitter). However, it is important to acknowledge that the data collection was conducted retrospectively. Twitter's moderation policies, which include censorship of spam messages, and the potential for users to delete their accounts or posts between the protests and the time of data collection, may affect the completeness of the dataset. Additionally, due to the contentious nature of the protests and the often ephemeral behaviour of social bots, there is an increased likelihood that bot-generated messages were deleted or accounts suspended, which could lead to an underestimation of bot activity compared to human users.

Nevertheless, the full Academic Track API has been discontinued[214], adding significant value to our dataset. With the API no longer available, our dataset represents a unique and comprehensive snapshot of protest-related discourse during a critical period. This discontinuation makes our data particularly valuable for historical and comparative analyses, as it captures a complete view of social media activity that is no longer accessible through current tools.

## Bot Identification

In our previous study[293], we developed a comprehensive methodology for bot identification on Twitter. This approach made use of the Botometer tool, which

calculates bot probabilities based on a multitude of features related to user profiles and activities. We applied specific thresholds (0.65 and 0.5) to classify accounts as 'automated,' 'human,' or 'unknown' based on these probabilities. In addition to Botometer, we trained several supervised machine learning models—including Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LOG), XGBoost (XGB), and Deep Learning (DL)—using a detailed set of features derived from various training datasets. These models were optimized to balance performance metrics such as specificity, sensitivity, and F1 score. To ensure robust bot detection, we combined the results from Botometer and our machine learning models through a triangulation approach. An account was classified as a bot if it was identified as such by both Botometer and at least one of our models.

Based on this methodology, we classified our dataset of all users active at the protest peak (k-core larger or equal to 2) into 'humans', 'bots' or 'unknown'.For a detailed description of the methodology and model performance, please refer to our earlier work[293].

We then classified the accounts into 'astroturfers' and 'other bots.' For identifying astroturfers, we applied a threshold of 2.5 out of 5 on the Botometer's display score. This threshold was chosen based on its alignment with our prior research, where it has proven effective in distinguishing astroturfing activities. A display score of 2.5/5 indicates that at least 50% of similar accounts in the Botometer' s training set are labeled as astroturfing bots[35]. This threshold provides a relatively reliable distinction between astroturfers and other bots. However, we acknowledge that this approach may introduce false positives. To address this concern, we conducted a series of robustness checks with thresholds ranging from 2.5 to 3.5 (see 'Model Tables'. Sample size for higher astroturfing score thresholds was insufficient for robust regression analysis). Despite variations in the threshold, the overall effect remained consistent.

## Opinion categorisation using ChatGPT

We used ChatGPT 3.5[218] to categorise the users' opinion on protests.The efficacy of leveraging large language models (LLMs) for these purposes has been widely explored in academic studies, and the methodology has been employed in prior research addressing similar textual classification tasks[219, 220]. We evaluated this classification technique in the context of protest-related social media data, within the Extinction Rebellion protests dataset. Their findings indicated that ChatGPT's classification achieved an accuracy rate of 0.88, compared to human coder assessments reported in our previous research[293]. This level of accuracy is generally regarded as both reliable and satisfactory.

The full prompt is as follows:

> On the text that I will give you in the following dialogues, please interpret it as a human with common political sense and background knowledge of racial justice issues. Tell me how positive do you think is this user's opinion towards the BLM protes. Give me a score that falls in the continuous range of -1 to 1. -1 being extremely negative, 1 extremely positive, 0 being neutral. If you think it is irrelevant to BLM, put 0 there. If the opinion seem to be mixed, put 0. Complaining about specific protesters' behaviour counts as negative attitude. Arguing that the protest's aim, goal or view of BLM is too extreme counts as a mild negative. Beware of sarcasm. People can support or do not support BLM protests with different partisan preferences, do not take that into account.Just return a score, no explanation needed. No full sentence needed either, the number itself will be fine. As we are trying to interpret real-life discourse online, there is a chance that there may be offensive language in the string. It is a virtual experiment evaluating people' s attitude, no one is actually hurt during the process. Here is the text:

## Topic modelling

We selected bi-term topic models because of their demonstrated success in extracting 'meaning'[99] and 'themes'[228]. This process is essential for understanding the context in which interactions (edges) are formed. Additionally, it enables us to analyze what each network cluster discusses by performing descriptive statistics on the topics.

Compared to traditional models such as Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI), bi-term models have shown greater accuracy in analysing short texts like tweets[228, 229]. We adopted a procedure of topic modelling similar to that we do for our previous paper, which also involves using topic modelling to extract theme of communication on short text on social media[293]. The procedure encompasses three main stages: 1) preprocessing the text data, 2) training and fine-tuning several topic modelling algorithms, and 3) applying the optimised model to the dataset.

During the preprocessing phase, we removed elements such as stopwords, usernames, emojis, and links from the tweets, and performed lemmatization and stemming on the remaining words. For stopwords, we used a standard list from the nltk package[231] and added Twitter-specific terms like 'RT', 'https', and 't.co'. Considering that retweets make up about 70% of tweets, we focused on unique tweets after preprocessing to more accurately capture prevailing trends.

To determine the optimal number of topics, we utilised both quantitative metrics and qualitative assessments. We experimented with bi-term topic models ranging from 2 to 15 topics. Quantitatively, we assessed the models using the U_mass coherence score, which evaluates the clarity and interpretability of the topics[233]. This metric helps gauge how comprehensible the topics generated by the model are[233].

From the U_mass coherence scores and manual assessments, we initially identified seven distinct topics within the corpus. Notably, Topic 1 was found to be too broad and was subsequently subdivided into five more specific sub-topics. This further classification was necessary because Topic 1 covered a wide range of concepts, making it less coherent. By breaking it down, we aim to achieve greater specificity and clarity, thus enhancing the interpretability and usefulness of the topics. Figure B.1 and B.2 shows the coherence scores for the two stages of topic modelling.

These topics were derived from computational models and were further refined through manual grouping based on thematic similarities. This approach allowed us

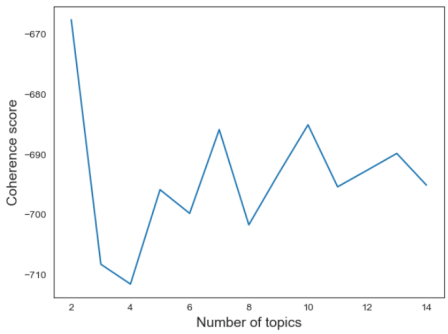



**Figure B.1:** Coherence scores of topic models (stage 1) with topic number from 2 to 15.

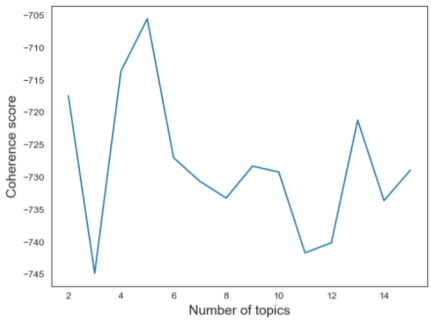



**Figure B.2:** Coherence scores of topic models (stage 2) with topic number from 2 to 15.

to consolidate the fragmented topics, which were highly focused on specific small-scale protest events across the United States, into four broader themes.

Organising topics into broader categories is a well-established practice in text analysis, as it helps in simplifying and interpreting complex datasets. Computer-assisted qualitative methodologies that interpret large quantity of text has effectively employed similar methods to manage and make sense of large corpora of text[375]. This approach facilitates a clearer understanding of the discourse[376], and consequentially enhances our analysis of opinion, information diffusion, and organisational patterns within communication networks.

In our study, we clustered the 15 initial topics into four overarching themes through a qualitative interpretation of the topic keywords. This consolidation was aimed at capturing the broader trends and dynamics in the communication networks more effectively. Table B.2 presents these four broad topics identified across the protest-related discourse during the peak period of the BLM protests.

**Table B.2:** Description of the four large topics with keywords and example posts

| Topics | Keywords | Sample Tweets |
|---|---|---|
| 1 - Protest live record and mobilization | blmnyc, pressfreedom, civiliberties, trader, habor, invokes, diagnosis, accelerate | 'Friday evening 640pm–Demo from St John the Devine pausing at 125th St #blmnyc' 'Second autopsy of George Floyd reveals unreported injuries and positive COVID-19 diagnosis #autopsy #BLM #report #BlackLivesMatter #Covid19 #GeorgeFloydWasMurdered' |
| 2 - Anti BLM conspiracies | candace, owens, hillary, coward, loser, trahshes, reich | 'Educate yourself Tony and Chelsea, Watch Candace Owens teachings. BLM, it was born after a lie in Ferguson MO. What I see is a bad cop and a suspect with a mile long record including pointing a gun at a pregnant woman's stomach.' |
| 3 - Spambots | dress, elpaso, daddy, [several interracial porn related hashtags redacted] | (Porn sellers using BLM hashtag - I believe it is better to not provide a sample.) |
| 4 - Kpop fans donating to BLM | army, bts, kpop, fandom, matchthemillion, donating, fan, armymatch, banbase | #ARMY, for those of you that can' t financially contribute to a BLM organization but would like to, I will be contributing $1 for every like and RT on this tweet through the #MatchAMillion campaign up to $300 USD. Borahae, ARMY! |

# Network construction and filtering

In this study, we constructed a bi-directional, unweighted retweet network (RT network) where nodes represent users and edges represent retweets or quotes between them. Edges were unweighted to reflect the presence of connections, simplifying the analysis by aggregating multiple interactions between node pairs into single edges. This approach helps to explore how interactions with bots might influence connections among human users.

'Retweets' are defined as posts shared with no additional comments, while 'quotes' include extra commentary. Each retweet or quote was recorded as a directional edge, with the original tweet poster as the starting node and the retweeter or commenter as the end node. Self-loops, indicating users who retweeted or quoted their own tweets, were included due to their impact on online discourse and mobilisation. Isolated nodes, representing users without original tweets or retweets, were removed.

For the analysis, we focused on a community of users who interacted with bots during a 48-hour snapshot of the BLM communication network. We used the k-core metric to filter the network, focusing on the maximal subgraph where each node is connected to at least k other nodes. We retained users with a k-core value

**Table B.3:** Descriptive statistics for k-core2 networks before, during and after protest peak

| | Before | Peak | After |
|---|---|---|---|
| Number of vertices (N) | 92078 | 103050 | 92078 |
| Number of edges | 740742 | 328708 | 913151 |
| Average unweighted degree | 16.09 | 6.38 | 19.83 |
| Maximum in-degree | 9155 | 12605 | 10747 |
| Maximum out-degree | 313 | 77 | 239 |
| Average clustering coefficient | 0.043 | 0.037 | 0.039 |
| Density | 0.000087 | 0.000031 | 0.000108 |
| Number of active bots | 39661 | 50982 | 38222 |
| Number of active humans | 17541 | 27221 | 17255 |

greater than two, resulting in a filtered network of 92,078 users and 372,932 edges. The filtering process was conducted using Python's NetworkX package and Gephi.

To measure the impact of bot interactions on network cohesiveness among human nodes, we constructed two additional network snapshots: the 'before' network (May 1, 2020, to May 2, 2020) and the 'after' network (September 1, 2020, to September 2, 2020). Directed communication networks were based on retweeting information, with full timeline data and user profiles collected for these periods.

In Table B.3 above, we present descriptive statistics for the network snapshots analysed, including the number of users, number of edges, and the average degree for each snapshot. It is important to note that the differences in the number of edges among the networks are attributable to the sampling methods used. The peak network, specifically, was sampled purely for analytical purposes and only included edges that are directly relevant to the protest. Our main focus is on understanding the changes between the 'before' and 'after' periods to evaluate the impact of the protests on network structure and bot interactions.

# Model Tables

**Table B.4:** Bot impact on individual level clustering (CAP=0.65, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.001 | −0.004* | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.010** | −0.010*** | | |
| | | | (0.004) | (0.003) | | |
| Other bot interactions at peak (logged) | | | | | 0.0002 | −0.002 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.531*** | | −0.530*** | | −0.530*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.007*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.016*** | | 0.015*** | | 0.015*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.023*** | | −0.021*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.003 | −0.106*** | 0.003** | −0.108*** | 0.002 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | −0.0001 | 0.443 | 0.0005 | 0.443 | −0.0001 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.5:** Bot impact on % of human-human edges for sub-communities (CAP=0.65, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.029* | −0.043*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.021 | −0.045** | | |
| | | | (0.017) | (0.016) | | |
| N other-human edges peak(logged) | | | | | −0.028* | −0.039** |
| | | | | | (0.013) | (0.012) |
| % of human-human edges before(humans) | | −0.901*** | | −0.903*** | | −0.879*** |
| | | (0.109) | | (0.112) | | (0.109) |
| Average ratio friends/followers (logged) | | −0.212 | | −0.192 | | −0.225 |
| | | (0.113) | | (0.117) | | (0.114) |
| Average ratio messages received/sent (logged) | | 0.018 | | −0.033 | | 0.017 |
| | | (0.072) | | (0.072) | | (0.073) |
| Average account age | | 0.041* | | 0.059** | | 0.040* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.122* | | −0.128* | | −0.129* |
| | | (0.060) | | (0.061) | | (0.060) |
| Median of favourite tweets (logged) | | 0.097 | | 0.099 | | 0.100 |
| | | (0.056) | | (0.057) | | (0.056) |
| Con-protest cluster (yes=1) | | −0.091 | | −0.081 | | −0.104 |
| | | (0.071) | | (0.074) | | (0.071) |
| Spam cluster (yes=1) | | −0.161* | | −0.106 | | −0.180* |
| | | (0.073) | | (0.076) | | (0.074) |
| Constant | −0.205*** | 0.092 | −0.263*** | −0.055 | −0.222*** | 0.128 |
| | (0.054) | (0.605) | (0.044) | (0.619) | (0.049) | (0.609) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.024 | 0.338 | 0.004 | 0.314 | 0.022 | 0.329 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.6:** Bot impact on individual level clustering, moderated by BLM support categories (CAP=0.65)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | 0.006 | 0.003 | | | | |
| | (0.004) | (0.003) | | | | |
| Total bot interactions at peak:Positive opinion | −0.008 | −0.009** | | | | |
| | (0.005) | (0.004) | | | | |
| Astroturfing bot interactions at peak (logged) | | | 0.010 | 0.006 | | |
| | | | (0.008) | (0.006) | | |
| Astroturfing bot interactions:Positive opinion | | | −0.025** | −0.020** | | |
| | | | (0.009) | (0.007) | | |
| Other bot interactions at peak (logged) | | | | | 0.003 | 0.002 |
| | | | | | (0.004) | (0.003) |
| Other bot interactions:Positive opinion | | | | | −0.003 | −0.006 |
| | | | | | (0.005) | (0.004) |
| Positive opinion to BLM (yes=1) | 0.001 | 0.005 | −0.001 | 0.001 | −0.003 | 0.002 |
| | (0.004) | (0.003) | (0.002) | (0.002) | (0.003) | (0.003) |
| Clustering coefficients before peak (humans) | | −0.531*** | | −0.530*** | | −0.530*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.007*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.017*** | | 0.015*** | | 0.016*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.024*** | | −0.024*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.002 | −0.111*** | 0.004 | −0.110*** | 0.003 | −0.108*** |
| | (0.003) | (0.010) | (0.002) | (0.010) | (0.003) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | 0.0002 | 0.443 | 0.001 | 0.444 | 0.00000 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.7:** Bot impact on % of human-human edges for sub-communities, moderated by BLM support categories (CAP=0.65)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.045 | −0.013 | | | | |
| | (0.044) | (0.031) | | | | |
| N astroturf-human edges peak(logged) | | | −0.036 | 0.008 | | |
| | | | (0.046) | (0.035) | | |
| N other-human edges peak(logged) | | | | | −0.033 | −0.007 |
| | | | | | (0.047) | (0.031) |
| Positive opinion to BLM (yes=1) | 0.030 | 0.206* | −0.005 | 0.149* | 0.077 | 0.198* |
| | (0.136) | (0.089) | (0.108) | (0.075) | (0.117) | (0.079) |
| N bot-human edges:positive opinion | 0.013 | −0.032 | | | | |
| | (0.047) | (0.032) | | | | |
| N astroturf-human edges:positive opinion | | | 0.016 | −0.054 | | |
| | | | (0.050) | (0.038) | | |
| N other-human edges:positive opinion | | | | | −0.0004 | −0.036 |
| | | | | | (0.049) | (0.033) |
| % of human-human edges before (humans) | | −0.749*** | | −0.731*** | | −0.729*** |
| | | (0.085) | | (0.088) | | (0.085) |
| Average account age | | 0.026* | | 0.047** | | 0.025 |
| | | (0.013) | | (0.014) | | (0.013) |
| Average ratio friends/followers (logged) | | −0.123 | | −0.112 | | −0.133 |
| | | (0.085) | | (0.089) | | (0.085) |
| Average ratio messages received/sent (logged) | | −0.0004 | | −0.049 | | −0.0002 |
| | | (0.054) | | (0.055) | | (0.054) |
| Median of statuses (logged) | | −0.054 | | −0.076 | | −0.059 |
| | | (0.047) | | (0.048) | | (0.047) |
| Median of favourite tweets (logged) | | 0.074 | | 0.079 | | 0.074 |
| | | (0.042) | | (0.043) | | (0.042) |
| Con-protest cluster (yes=1) | | −0.063 | | −0.049 | | −0.078 |
| | | (0.053) | | (0.055) | | (0.053) |
| Spam cluster (yes=1) | | −0.045 | | −0.067 | | −0.059 |
| | | (0.076) | | (0.081) | | (0.073) |
| Constant | −0.207 | 0.802 | −0.255** | 0.816 | −0.257** | 0.859 |
| | (0.118) | (0.488) | (0.095) | (0.504) | (0.098) | (0.485) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.017 | 0.360 | −0.008 | 0.323 | 0.017 | 0.352 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table B.8:** Bot impact on individual level clustering (CAP=0.50, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.002 | −0.005* | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.010** | −0.010*** | | |
| | | | (0.004) | (0.003) | | |
| Other bot interactions at peak (logged) | | | | | −0.001 | −0.002 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.530*** | | −0.530*** | | −0.530*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.007*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.017*** | | 0.015*** | | 0.015*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.022*** | | −0.021*** | | −0.022*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.004* | −0.107*** | 0.003** | −0.108*** | 0.003 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | 0.00003 | 0.443 | 0.0005 | 0.443 | −0.00004 | 0.443 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table B.9:** Bot impact on % of human-human edges for sub-communities (CAP=0.50, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.038** | −0.045*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.021 | −0.045** | | |
| | | | (0.017) | (0.016) | | |
| N other-human edges peak(logged) | | | | | −0.036** | −0.041*** |
| | | | | | (0.013) | (0.012) |
| % of human-human edges before(humans) | | −0.870*** | | −0.903*** | | −0.854*** |
| | | (0.108) | | (0.112) | | (0.108) |
| Average ratio friends/followers (logged) | | −0.217 | | −0.192 | | −0.228* |
| | | (0.113) | | (0.117) | | (0.113) |
| Average ratio messages received/sent (logged) | | 0.019 | | −0.033 | | 0.019 |
| | | (0.072) | | (0.072) | | (0.072) |
| Average account age | | 0.040* | | 0.059** | | 0.040* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.124* | | −0.128* | | −0.129* |
| | | (0.060) | | (0.061) | | (0.060) |
| Median of favourite tweets (logged) | | 0.106 | | 0.099 | | 0.108 |
| | | (0.056) | | (0.057) | | (0.056) |
| Con-protest cluster (yes=1) | | −0.084 | | −0.081 | | −0.093 |
| | | (0.071) | | (0.074) | | (0.071) |
| Spam cluster (yes=1) | | −0.156* | | −0.106 | | −0.167* |
| | | (0.072) | | (0.076) | | (0.073) |
| Constant | −0.166** | 0.042 | −0.263*** | −0.055 | −0.181*** | 0.065 |
| | (0.058) | (0.603) | (0.044) | (0.619) | (0.054) | (0.607) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.042 | 0.342 | 0.004 | 0.314 | 0.040 | 0.335 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table B.10:** Bot impact on individual level clustering (CAP=0.65, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.001 | −0.004* | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.010** | −0.010*** | | |
| | | | (0.004) | (0.003) | | |
| Other bot interactions at peak (logged) | | | | | −0.001 | −0.002 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.530*** | | −0.530*** | | −0.530*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.007*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.016*** | | 0.015*** | | 0.015*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.022*** | | −0.021*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.003 | −0.106*** | 0.003** | −0.108*** | 0.002 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | −0.00003 | 0.443 | 0.0005 | 0.443 | −0.0001 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.11:** Bot impact on % of human-human edges for sub-communities (CAP=0.55, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.034* | −0.042*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.021 | −0.045** | | |
| | | | (0.017) | (0.016) | | |
| N other-human edges peak(logged) | | | | | −0.033* | −0.039** |
| | | | | | (0.014) | (0.012) |
| % of human-human edges before(humans) | | −0.873*** | | −0.903*** | | −0.857*** |
| | | (0.109) | | (0.112) | | (0.109) |
| Average ratio friends/followers (logged) | | −0.219 | | −0.192 | | −0.230* |
| | | (0.114) | | (0.117) | | (0.114) |
| Average ratio messages received/sent (logged) | | 0.018 | | −0.033 | | 0.018 |
| | | (0.072) | | (0.072) | | (0.073) |
| Average account age | | 0.041* | | 0.059** | | 0.040* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.125* | | −0.128* | | −0.130* |
| | | (0.060) | | (0.061) | | (0.061) |
| Median of favourite tweets (logged) | | 0.103 | | 0.099 | | 0.104 |
| | | (0.056) | | (0.057) | | (0.056) |
| Con-protest cluster (yes=1) | | −0.090 | | −0.081 | | −0.100 |
| | | (0.072) | | (0.074) | | (0.072) |
| Spam cluster (yes=1) | | −0.155* | | −0.106 | | −0.167* |
| | | (0.073) | | (0.076) | | (0.073) |
| Constant | −0.182** | 0.066 | −0.263*** | −0.055 | −0.197*** | 0.092 |
| | (0.057) | (0.607) | (0.044) | (0.619) | (0.053) | (0.610) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.033 | 0.334 | 0.004 | 0.314 | 0.031 | 0.327 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table B.12:** Bot impact on individual level clustering (CAP=0.60, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.002 | −0.004* | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.010** | −0.010*** | | |
| | | | (0.004) | (0.003) | | |
| Other bot interactions at peak (logged) | | | | | −0.001 | −0.002 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.530*** | | −0.530*** | | −0.530*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.007*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.016*** | | 0.015*** | | 0.015*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.022*** | | −0.021*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.003 | −0.106*** | 0.003** | −0.108*** | 0.003 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | −0.00001 | 0.443 | 0.0005 | 0.443 | −0.0001 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.13:** Bot impact on % of human-human edges for sub-communities (CAP=0.60, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.034* | −0.043*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.021 | −0.045** | | |
| | | | (0.017) | (0.016) | | |
| N other-human edges peak(logged) | | | | | −0.032* | −0.039*** |
| | | | | | (0.013) | (0.012) |
| % of human-human edges before(humans) | | −0.878*** | | −0.903*** | | −0.861*** |
| | | (0.108) | | (0.112) | | (0.109) |
| Average ratio friends/followers (logged) | | −0.217 | | −0.192 | | −0.228* |
| | | (0.113) | | (0.117) | | (0.114) |
| Average ratio messages received/sent (logged) | | 0.019 | | −0.033 | | 0.019 |
| | | (0.072) | | (0.072) | | (0.073) |
| Average account age | | 0.041* | | 0.059** | | 0.040* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.125* | | −0.128* | | −0.131* |
| | | (0.060) | | (0.061) | | (0.060) |
| Median of favourite tweets (logged) | | 0.100 | | 0.099 | | 0.102 |
| | | (0.056) | | (0.057) | | (0.056) |
| Con-protest cluster (yes=1) | | −0.090 | | −0.081 | | −0.101 |
| | | (0.071) | | (0.074) | | (0.072) |
| Spam cluster (yes=1) | | −0.159* | | −0.106 | | −0.173* |
| | | (0.073) | | (0.076) | | (0.073) |
| Constant | −0.186** | 0.105 | −0.263*** | −0.055 | −0.201*** | 0.133 |
| | (0.056) | (0.605) | (0.044) | (0.619) | (0.051) | (0.609) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.033 | 0.337 | 0.004 | 0.314 | 0.031 | 0.330 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.14:** Bot impact on individual level clustering (CAP=0.70, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.001 | −0.006** | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.012** | −0.012*** | | |
| | | | (0.004) | (0.003) | | |
| Other bot interactions at peak (logged) | | | | | −0.0002 | −0.003 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.531*** | | −0.530*** | | −0.531*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.008*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001 | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.017*** | | 0.015*** | | 0.016*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.023*** | | −0.021*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.003 | −0.106*** | 0.003** | −0.108*** | 0.002 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | −0.00004 | 0.443 | 0.001 | 0.443 | −0.0001 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.15:** Bot impact on % of human-human edges for sub-communities (CAP=0.70, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.028* | −0.041*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.021 | −0.045** | | |
| | | | (0.017) | (0.016) | | |
| N other-human edges peak(logged) | | | | | −0.027 | −0.038** |
| | | | | | (0.014) | (0.012) |
| % of human-human edges before(humans) | | −0.899*** | | −0.903*** | | −0.883*** |
| | | (0.109) | | (0.112) | | (0.111) |
| Average ratio friends/followers (logged) | | −0.218 | | −0.192 | | −0.225 |
| | | (0.114) | | (0.117) | | (0.116) |
| Average ratio messages received/sent (logged) | | 0.014 | | −0.033 | | 0.009 |
| | | (0.072) | | (0.072) | | (0.074) |
| Average account age | | 0.041* | | 0.059** | | 0.040* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.123* | | −0.128* | | −0.129* |
| | | (0.060) | | (0.061) | | (0.062) |
| Median of favourite tweets (logged) | | 0.099 | | 0.099 | | 0.103 |
| | | (0.056) | | (0.057) | | (0.058) |
| Con-protest cluster (yes=1) | | −0.092 | | −0.081 | | −0.110 |
| | | (0.071) | | (0.074) | | (0.074) |
| Spam cluster (yes=1) | | −0.163* | | −0.106 | | −0.187* |
| | | (0.073) | | (0.076) | | (0.078) |
| Constant | −0.214*** | 0.079 | −0.263*** | −0.055 | −0.226*** | 0.084 |
| | (0.053) | (0.607) | (0.044) | (0.619) | (0.049) | (0.621) |
| Observations | 150 | 150 | 150 | 150 | 146 | 146 |
| Adjusted $R^2$ | 0.021 | 0.334 | 0.004 | 0.314 | 0.020 | 0.326 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table B.16:** Bot impact on individual level clustering (CAP=0.75, Astroturfing threshold=2.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.004 | −0.008*** | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.015*** | −0.011** | | |
| | | | (0.005) | (0.003) | | |
| Other bot interactions at peak (logged) | | | | | −0.002 | −0.006*** |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.531*** | | −0.530*** | | −0.531*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.009*** | | −0.007*** | | −0.008*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001 | | 0.001* | | 0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.018*** | | 0.015*** | | 0.017*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.023*** | | −0.021*** | | −0.024*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.004** | −0.106*** | 0.003** | −0.107*** | 0.003* | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | 0.0002 | 0.444 | 0.001 | 0.443 | 0.00003 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.17:** Bot impact on individual level clustering (CAP=0.65, Astroturfing threshold=3)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.001 | −0.004* | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.014** | −0.012*** | | |
| | | | (0.005) | (0.004) | | |
| Other bot interactions at peak (logged) | | | | | 0.0003 | −0.003 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.531*** | | −0.530*** | | −0.530*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.007*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.016*** | | 0.015*** | | 0.015*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.023*** | | −0.021*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.003 | −0.106*** | 0.003** | −0.108*** | 0.002 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | −0.0001 | 0.443 | 0.001 | 0.443 | −0.0001 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.18:** Bot impact on % of human-human edges for sub-communities (CAP=0.65, Astroturfing threshold=3)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.029* | −0.043*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.026 | −0.045** | | |
| | | | (0.018) | (0.017) | | |
| N other-human edges peak(logged) | | | | | −0.028* | −0.041*** |
| | | | | | (0.014) | (0.012) |
| % of human-human edges before(humans) | | −0.901*** | | −0.890*** | | −0.889*** |
| | | (0.109) | | (0.112) | | (0.109) |
| Average ratio friends/followers (logged) | | −0.212 | | −0.198 | | −0.220 |
| | | (0.113) | | (0.117) | | (0.114) |
| Average ratio messages received/sent (logged) | | 0.018 | | −0.028 | | 0.015 |
| | | (0.072) | | (0.072) | | (0.072) |
| Average account age | | 0.041* | | 0.056** | | 0.041* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.122* | | −0.126* | | −0.129* |
| | | (0.060) | | (0.061) | | (0.060) |
| Median of favourite tweets (logged) | | 0.097 | | 0.098 | | 0.101 |
| | | (0.056) | | (0.057) | | (0.056) |
| Con-protest cluster (yes=1) | | −0.091 | | −0.088 | | −0.098 |
| | | (0.071) | | (0.073) | | (0.071) |
| Spam cluster (yes=1) | | −0.161* | | −0.108 | | −0.176* |
| | | (0.073) | | (0.076) | | (0.073) |
| Constant | −0.205*** | 0.092 | −0.262*** | −0.053 | −0.216*** | 0.113 |
| | (0.054) | (0.605) | (0.041) | (0.621) | (0.051) | (0.607) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.024 | 0.338 | 0.007 | 0.311 | 0.022 | 0.333 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

**Table B.19:** Bot impact on individual level clustering (CAP=0.65, Astroturfing threshold=3.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Change in clustering for ego networks before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total bot interactions at peak (logged) | −0.001 | −0.004* | | | | |
| | (0.002) | (0.002) | | | | |
| Astroturfing bot interactions at peak (logged) | | | −0.023*** | −0.013** | | |
| | | | (0.007) | (0.005) | | |
| Other bot interactions at peak (logged) | | | | | −0.00001 | −0.004 |
| | | | | | (0.002) | (0.002) |
| Clustering coefficients before peak (humans) | | −0.531*** | | −0.530*** | | −0.531*** |
| | | (0.005) | | (0.005) | | (0.005) |
| Ratio between friends and followers (logged) | | −0.001 | | −0.001 | | −0.001 |
| | | (0.001) | | (0.001) | | (0.001) |
| Ratio between messages received and sent (logged) | | −0.008*** | | −0.007*** | | −0.008*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Account age | | −0.003*** | | −0.003*** | | −0.003*** |
| | | (0.0002) | | (0.0002) | | (0.0002) |
| Number of statuses (logged) | | 0.012*** | | 0.012*** | | 0.012*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Number of favourite tweets (logged) | | 0.001* | | 0.001* | | 0.001* |
| | | (0.001) | | (0.001) | | (0.001) |
| Total number of degrees at protest peak (logged) | | 0.016*** | | 0.015*** | | 0.016*** |
| | | (0.002) | | (0.001) | | (0.002) |
| Con-protest cluster (yes=1) | | −0.019*** | | −0.019*** | | −0.019*** |
| | | (0.003) | | (0.003) | | (0.003) |
| Spam cluster (yes=1) | | −0.023*** | | −0.022*** | | −0.023*** |
| | | (0.006) | | (0.006) | | (0.006) |
| Constant | 0.003 | −0.106*** | 0.003** | −0.107*** | 0.002 | −0.106*** |
| | (0.002) | (0.010) | (0.001) | (0.010) | (0.002) | (0.010) |
| Observations | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 | 12,921 |
| Adjusted $R^2$ | −0.0001 | 0.443 | 0.001 | 0.443 | −0.0001 | 0.443 |

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table B.20:** Bot impact on % of human-human edges for sub-communities (CAP=0.65, Astroturfing threshold=3.5)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Ratio change in % of human-human edges before and after protest peak (Scaled) | | | | | |
| | Total | Total | Astroturfing bots | Astroturfing bots | Other bots | Other bots |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N bot-human edges peak(logged) | −0.029* | −0.043*** | | | | |
| | (0.014) | (0.012) | | | | |
| N astroturf-human edges peak(logged) | | | −0.025 | −0.044* | | |
| | | | (0.020) | (0.018) | | |
| N other-human edges peak(logged) | | | | | −0.029* | −0.041*** |
| | | | | | (0.014) | (0.012) |
| % of human-human edges before(humans) | | −0.901*** | | −0.884*** | | −0.891*** |
| | | (0.109) | | (0.112) | | (0.109) |
| Average ratio friends/followers (logged) | | −0.212 | | −0.203 | | −0.220 |
| | | (0.113) | | (0.117) | | (0.114) |
| Average ratio messages received/sent (logged) | | 0.018 | | −0.034 | | 0.017 |
| | | (0.072) | | (0.072) | | (0.072) |
| Average account age | | 0.041* | | 0.054** | | 0.041* |
| | | (0.018) | | (0.019) | | (0.018) |
| Median of statuses (logged) | | −0.122* | | −0.130* | | −0.126* |
| | | (0.060) | | (0.062) | | (0.060) |
| Median of favourite tweets (logged) | | 0.097 | | 0.100 | | 0.099 |
| | | (0.056) | | (0.057) | | (0.056) |
| Con-protest cluster (yes=1) | | −0.091 | | −0.088 | | −0.095 |
| | | (0.071) | | (0.074) | | (0.071) |
| Spam cluster (yes=1) | | −0.161* | | −0.120 | | −0.166* |
| | | (0.073) | | (0.075) | | (0.073) |
| Constant | −0.205*** | 0.092 | −0.271*** | −0.023 | −0.211*** | 0.107 |
| | (0.054) | (0.605) | (0.039) | (0.623) | (0.052) | (0.606) |
| Observations | 150 | 150 | 150 | 150 | 150 | 150 |
| Adjusted $R^2$ | 0.024 | 0.338 | 0.004 | 0.305 | 0.023 | 0.335 |
| *Note:* | | | | | | *p<0.05; **p<0.01; ***p<0.001 |

# C

# Supplementary information for Chapter 6

This appendix provides supplementary information for Chapter 6 of this dissertation. It details the methodological procedures and rationale for method selection for the following aspects: data collection, bot identification, data sampling, Computer-assisted content analysis and support group categorisation.

## Case selection

This research used two representative protest cases for our content analysis: XR and BLM. The selection of these cases is based on several factors, including the nature and objectives of the protests, and the feasibility of conducting CACA to achieve the research objectives. Both XR and BLM are globally recognised protests that have attracted significant attention, with activists actively engaging on various social media platforms, particularly X (formerly Twitter). Social media platforms serve as channels for information dissemination, resonance, and gaining support in those cases[38, 377], making them suitable cases for analysing factors influencing activists' discourse opportunity structure. The extensive discussions and abundant data generated during these events is particularly suitable for in-depth analysis of the DOS online. Additionally, previous studies and news reports have indicated the presence of social bots in discussions related to climate change and racial issues

on social media, further justifying the relevance of studying the role of bots in these specific contexts[296, 43].

XR, or Extinction Rebellion, is a grassroots global environmental social movement that emerged in the UK in October 2018 and has since maintained its presence. It operates as a decentralized movement, with activists organizing autonomously in line with XR's principles and goals. The movement employs non-violent disobedience tactics to draw attention to ecological destruction and climate change, advocating for governmental action[176]. XR activists have organised numerous disruptive events across the UK and internationally. During the process they relied heavily on social media platforms like Facebook, Instagram, and X for communication and outreach[45]. The analysis focuses on discourse surrounding a series of protest events that occurred from November to December 2019, which is a period of significant XR activism[378, 176].

The other case examined is the Black Lives Matter (BLM) movement, which originated in 2014 in Ferguson, Missouri, following the death of Michael Brown. BLM aims to raise awareness about police violence against unarmed African Americans and combat racism and racial inequality[191]. Similar to XR, BLM activists heavily rely on social media, especially X, to disseminate messages and mobilise support[39]. The analysis focuses on the protest events from late May to early June 2020 following the death of George Floyd. Floyd, an African American man, died while in police custody, sparking widespread outrage and protests across the United States[379]. This case represents another instance of large-scale protests using social media platforms to mobilise and disseminate information.

## Data collection

This study makes use of a sample of bot activities from two distinct datasets to analyse X (formerly Twitter) activity around the two protest events. The first dataset contains communication during the Extinction Rebellion climate change protests from 18 November, 2019, to 10 December, 2019. Data was collected between November and December 2022 using Twitter's Academic Research product

track API, which provided a comprehensive historical archive of tweets[213] and was available by the time of data collection. We focused on tweets containing keywords including 'Extinction Rebellion', 'climate change protest', 'XRebellion', 'XR', and various keyword variants. The dataset encompasses 201,010 tweets from 122,130 users.

The second dataset examines activity on X related to the Black Lives Matter (BLM) protests following George Floyd' s death, with a focus on the peak protest period from 6 June to 8 June, 2020[379]. Using the same Twitter Academic Product track API, we collected tweets containing keywords like 'BLM', 'George Floyd, and their variants. This collection includes over 2.5 million tweets from more than one million users between 6 June, 2020, 00:00, and 8 June, 2020, 00:00.

For this study, only a small proportion of these tweets was selected for detailed analysis, along with user data retrieved via the Twitter user lookup API (available by the time of data collection)[380]. The procedure of further filtering is detailed in the 'data sampling' section below.

Retrospective data collection presents some limitations, such as potential underestimation of social bot activity due to Twitter' s spam censorship and user account changes. Additionally, while data was collected after the protest events, we assume that the number of 'likes' on tweets remains relatively stable post-event due to the focused and time-sensitive nature of the discourse.

## Bot identification

In our prior research[293], we established a robust methodology for identifying bots on Twitter, which is applied here as a foundational reference. We utilised the Botometer tool[35], known for its machine learning-based classification of bot likelihood, assessing various features related to user profiles and interactions. We set thresholds of 0.65 and 0.5 to categorise accounts into 'automated,' 'human,' or 'unknown' based on the complete automation probability (CAP) scores.

Alongside Botometer, we developed and trained several supervised machine learning models—Random Forest (RF), Support Vector Machine (SVM), Logistic

Regression (LOG), XGBoost (XGB), and Deep Learning (DL)—using a comprehensive set of features from training datasets containing both bots and human accounts. These models were fine-tuned to optimise performance across several metrics, including specificity, sensitivity, and F1 score.

To enhance the accuracy of bot detection, we integrated results from both the Botometer tool and our machine learning models through a triangulation strategy. Accounts were classified as bots if they were identified as such by both Botometer and at least one of our models.

For further details on the methodology and model performance, please refer to our previous publication[293]. The method is then adopted on the two protest users dataset mentioned above for further classification and categorisation tasks.

## Data sampling

The procedure for filtering the XR1, XR2, BLM1, and BLM2 datasets involved multiple steps to ensure data relevance and accuracy. First, tweets from users categorised as 'bots' were filtered, with tweets further categorised as either retweets (XR1, BLM1) or original posts (XR2, BLM2) to effectively capture bot activity. Given the large sample size and significant proportion of bot users (55.3% for BLM and 48% for XR), text was standardised by lower-casing, removing punctuation, and lemmatising. Duplicates were then removed from the datasets while retaining the count for each distinct tweet. This process resulted in 2,513 posts for XR1, 567 for XR2, 5,383 for BLM1, and 1,438 for BLM2, along with information on the number of retweets by bots and humans, and duplicated original messages.

Additionally, the corpus includes two supplementary datasets. The first, USR1, was created by filtering tweets that discuss social bots using keywords such as 'bot,' 'social bot,' and 'automated accounts' posted by human users during peak protest periods. This dataset includes 467 posts discussing social bots during XR and BLM protests, providing insight into how human activists perceive and react to bot activity. The second supplementary dataset, RD1, contains social media posts from Reddit. RD1 was collected using keyword retrieval methods, ensuring the

dataset's relevance by filtering out irrelevant posts. Reddit's post-reply structure allows for more in-depth discussions focused on specific bot features and strategies, aiding in the identification and analysis of new bot tactics.

## Computer-assisted content analysis

The Computer-assisted content analysis (CACA) approach integrates computational tools with traditional content analysis to systematically analyse large textual datasets[358]. In CACA methods, NLP algorithms are employed to process and analyse extensive datasets, so as to identify patterns and insights that would be impractical to discern manually[381, 359].

This method is particularly effective for analysing the impact of bots on the DOS of activists because it combines computational efficiency with detailed pattern recognition. By employing NLP algorithms, CACA can process and analyse extensive datasets at scale, which is important for capturing the complex and often subtle ways in which bots influence online discourse. This method allows for the automated detection of patterns in textual interactions, enabling researchers to identify how bots affect key DOS dimensions—such as the visibility of certain topics, the resonance of specific messages, and the overall legitimacy of the discourse. Traditional manual analysis would struggle to manage the volume and complexity of data involved, but CACA ' s computational tools provide a systematic and comprehensive approach to uncovering these impacts, offering deeper insights into how bots shape and alter social media conversations.

## ChatGPT assisted text coding procedures

We used GPT[218] to perform a classification of the text following the framework of social bots we developed online. For each bot-generated message, we provided the model with an instruction prompt on how to classify the content based on predetermined categories, namely dimensions and sub-dimensions of bot strategies of

DOS that we discovered, along with the text to classify (users' tweets). The model then outputted the corresponding code.

This method was adopted in previous research that involves similar textual classification tasks[382, 219]. The validity of using LLMs for such tasks has been discussed extensively in the literature[383, 384]. Specifically, we[293] examined the effectiveness of this classification approach in the context of protests using the same dataset of social media records in XR protests. ChatGPT classification achieved an accuracy of 0.88 compared to the judgement of human coder in our past research[293].

The prompts used for text coding tasks as follows:

## Visibility

> Interpret the following strings with an understanding of modern democratic politics and common political sense. Categorise the function of each tweet in the XR or BLM movement discussions using these labels:
>
> 1.Pro-protest: Supports protests, spreads protest-related information, or aids protest mobilisation.
>
> 2.Con-protest: Opposes protests, protest mobilisation, or the relevant movement.
>
> 3.News: Appears to be a neutral news headline. If there' s any apparent sentiment, categorise as either 'pro-protest' or 'con-protest.'
>
> 4.Irrelevant: Unrelated to the protest. This includes content only tagged with #blm or #xr but otherwise irrelevant.
>
> 5.Other political topics: Mentions XR or BLM but mainly focuses on other political issues (e.g., criticising a political figure).

## Resonance

> Please interpret the following tweet with common political sense and background knowledge of modern democratic politics and XR/BLM. Determine its function in the XR/BLM movement discussions using three dimensions:
>
> 1. Categorise the tweet as 'positive' or 'negative' to XR/BLM activism. If neutral, put 'neutral'. Return only one category in this dimension.
>
> 2. Categorise the tweet by message type: 'Direct Provocation': use of statements to elicit an immediate and strong emotional response. Examples include posting inflammatory or offensive remarks (not supportive) or outright uplifting or motivational statements (supportive).

'Amplified Neutrality': amplifying a neutral message in a context that heightens emotional impact. It involves emphasising a neutral statement or neutral message in a positive light or inflammatory context. If neither of these two fit, put 'not applicable'. Categories do not overlap, they are either one category or not applicable.

3. Categorise the tweet by topic: 'cultural resonance': if message aim to evoke a response based on shared cultural values, norms, and identities. 'emotional resonance': if message aim to provoke an emotional reaction on an individual level. If neither of these two fit, put 'not applicable'. Categories can overlap, they are either one category, two categories or not applicable.

Return categories separated by commas, in the format of 'helpful, direct provocation, cultural resonance'. Here is the string:

## Legitimacy

Please interpret the following tweet with common political sense and background knowledge of modern democratic politics. Determine its function in the XR/BLM movement discussions using two dimensions:

1. Categorise the tweet as 'helpful' or 'unhelpful' to XR/BLM activism. If neutral, put 'neutral'. Return only one category in this dimension.

2. Categorise the tweet by topic: 'legitimacy of protest frame' (aims/goals of XR/BLM), 'image of activists' (behaviour of activists/supporters or anti-protest activists), 'protest participation' (information or encouragement/discouragement about protest activity), 'trustworthiness of the user' (author's credibility), or 'not applicable'. Categories can overlap; return all applicable ones separated by commas.

# D
# Authorship statement

The research presented in Chapter 3 (Methodology) and Chapter 4 (The impact of social bots to human behaviour in protest communication) was conducted in collaboration with Dr. Orsolya Vasarhelyi from the Centre for Collective Learning and the Institute of Data Analytics and Information Systems at Corvinus University, and Prof. Balazs Vedres from Central European University. A journal article based on the materials in the two chapters is accepted for publication at *Scientific Reports*.

I, Linda Li, am the lead author of the submitted paper. I was responsible for the original conceptualisation of the research question and the design of the study. I also gathered all the research data, developed the methodological framework, performed the majority of the analysis, and drafted the manuscript. OV contributed by aiding with refining my methods for social bot detection and text classification using ChatGPT, particularly in the work discussed in Chapter 3, with her expertise in machine learning. For the analysis of social bots' impact on communication (Chapter 4), both OV and BV provided guidance on framing the work. They also helped replicate and refine the statistical analysis, conducted robustness checks, and assisted in improving the visualisations and manuscript.

The current versions of all other chapters, as submitted for the dissertation, are entirely the work of Linda Li.

(Attached: signed copy of the statement by all co-authors)

# Authorship Statement

The research presented in Chapter 3 and Chapter 4 on the impact of social bots on human behaviour in political communication was conducted in collaboration with Dr. Orsolya Vasarhelyi from the Center for Collective Learning and the Institute of Data Analytics and Information Systems at Corvinus University, and Prof. Balazs Vedres from Central European University. This chapter is currently accepted for publication in *Scientific Reports.*

I, Linda Li, am the lead author of the submitted article. I was responsible for the original conceptualisation of the research question and the design of the study. I also gathered all the research data, developed the overall methodological framework, performed the analysis predominately, and drafted the manuscript.

OV contributed by aiding with refining my methods for social bot detection and text classification using ChatGPT, particularly in the work discussed in Chapter 3, with her expertise in machine learning. For the analysis of social bots' impact on communication (Chapter 4), both OV and BV provided guidance on framing the work. They also helped replicate and refine the statistical analysis, conducted robustness checks, and assisted in improving the visualisations and manuscript.

The current versions of all other chapters, as submitted for the dissertation, are entirely the work of Linda Li.

Name: Linda Li

Date:20/09/2024

Signature: 

I, Orsolya Vasarhelyi, certify that the above is an accurate description of the candidate's contribution as lead author of the article for which I am listed as co-author.

Name: Orsolya Vasarhelyi

Date: 20-09-2024

Signature: 

I, Balazs Vedres, certify that the above is an accurate description of the candidate's contribution as lead author of the article for which I am listed as co-author.

Name: Balazs Vedres

Date: 23-09-2024



# Bibliography


[1] Evan Stewart and Douglas Hartmann. The New Structural Transformation of the Public Sphere. *Sociological Theory*, 38(2):170–191, 2020.

[2] Andreas Hepp. Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies. *Media, Culture and Society*, 42(7-8):1410–1426, 2020.

[3] Joshua A. Tucker, Yannis Theocharis, Margaret E. Roberts, and Pablo Barberá. From liberation to turmoil: Social media and democracy. *Journal of Democracy*, 28(4):46–59, 2017.

[4] William Lafi Youmans and Jillian C. York. Social Media and the Activist Toolkit: User Agreements, Corporate Interests, and the Information Infrastructure of Modern Social Movements. *Journal of Communication*, 62(2):315–329, 2012.

[5] Neal Caren, Kenneth T. Andrews, and Todd Lu. Contemporary social movements in a hybrid media environment. *Annual Review of Sociology*, 46:443–465, 2020.

[6] Dustin Kidd and Keith McIntosh. Social media and social movements. *Sociology Compass*, 10(9):785–794, 2016.

[7] Jeroen Van Laer and Peter Van Aelst. Internet and social movement action repertoires: Opportunities and limitations. *Information, communication & society*, 13(8):1146–1171, 2010.

[8] W. Lance Bennett and Alexandra Segerberg. The logic of connective action: Digital media and the personalization of contentious politics. *Information Communication and Society*, 15(5):739–768, 2012.

[9] Xie Tang and Ziyi Xu. How Does Digital Technology Change Climate Activism? Take Extinction Rebellion as A Case. *Journal of Social Science and Humanities*, 4(10):29–33, 2022.

[10] Joshua Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*, 2018.

[11] Daniel Kreiss and Shannon C McGregor. A review and provocation: On polarization and platforms. *New Media & Society*, 26(1):556–579, 2024.

[12] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[13] Marco T. Bastos and Dan Mercea. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, 37(1):38–54, 2019.

[14] Tobias R. Keller and Ulrike Klinger. Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications. *Political Communication*, 36(1):171–189, 2019.

[15] Dennis Assenmacher, Lena Clever, Lena Frischlich, Thorsten Quandt, Heike Trautmann, and Christian Grimme. Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media and Society*, 6(3), 2020.

[16] Catlin Dewey. The next frontier of online activism is woke chatbots. https://www.washingtonpost.com/news/the-intersect/wp/2016/08/11/the-next-frontier-of-online-activism-is-woke-chatbots/, 2016. Accessed: 2024-10-09.

[17] Ho-Chun Herbert Chang, Emily Chen, Meiqing Zhang, Goran Muric, and Emilio Ferrara. Social bots and social media manipulation in 2020: The year in review. In *Handbook of Computational Social Science, Volume 1*, pages 304–323. Routledge, 2021.

[18] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences of the United States of America*, 115(49):12435–12440, 2018.

[19] Stefano Cresci. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83, 2020.

[20] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social Bots: Human-Like by Means of Human Control? *Big Data*, 5(4):279–293, 2017.

[21] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[22] Yini Zhang, Dhavan Shah, Jordan Foley, Aman Abhishek, Josephine Lukito, Jiyoun Suk, Sang Jung Kim, Zhongkai Sun, Jon Pevehouse, and Christine Garlough. Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication*, 24(4):182–202, 2019.

[23] Sandra González-Bailón and Manlio De Domenico. Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences of the United States of America*, 118(11), 2021.

[24] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1), 2018.

[25] Deen Freelon, Charlton McIlwain, and Meredith Clark. Quantifying the power and consequences of social media protest. *New Media and Society*, 20(3):990–1011, 2018.

[26] Will Jennings and Clare Saunders. Street Demonstrations and the Media Agenda: An Analysis of the Dynamics of Protest Agenda Setting. *Comparative Political Studies*, 52(13-14):2283–2313, 2019.

[27] Marco T. Bastos, Dan Mercea, and Arthur Charpentier. Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, 65(2):320–350, 2015.

[28] John T Jost, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A Tucker. How social media facilitates political protest: Information, motivation, and social networks. *Political psychology*, 39:85–118, 2018.

[29] Bruce Bimber. Three Prompts for Collective Action in the Context of Digital Media. *Political Communication*, 34(1):6–20, 2017.

[30] Tuuli Marja Kleiner. Public opinion polarisation and protest behaviour. *European Journal of Political Research*, 57(4):941–962, 2018.

[31] Alicia A. Weaver. Does Protest Behavior Mediate the Effects of Public Opinion on National Environmental Policies? A Simple Question and a Complex Answer. *International Journal of Sociology*, 38(3):108–125, 2008.

[32] Todd Rogers, Noah J Goldstein, and Craig R Fox. Social mobilization. *Annual review of psychology*, 69:357–381, 2018.

[33] Michelle Forelle, Phil Howard, Andrés Monroy-Hernández, and Saiph Savage. Political bots and the manipulation of public opinion in venezuela, 2015.

[34] Loni Hagen, Stephen Neely, Thomas E Keller, Ryan Scharf, and Fatima Espinoza Vasquez. Rise of the machines? examining the influence of social bots on a political discussion network. *Social Science Computer Review*, 40(2):264–287, 2022.

[35] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. Botometer 101: Social bot practicum for computational social scientists. *Journal of computational social science*, 5(2):1511–1528, 2022.

[36] Adrian Rauchfleisch and Jonas Kaiser. The false positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10):1–20, 10 2020.

[37] Christian Grimme, Dennis Assenmacher, and Lena Adam. Changing Perspectives: Is It Sufficient to Detect Social Bots? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10913 LNCS(June):445–461, 2018.

[38] Peter Gardner, Tiago Carvalho, and Maria Valenstain. Spreading rebellion?: The rise of extinction rebellion chapters across the world. *Environmental Sociology*, 8(4):424–435, 2022.

[39] Jelani Ince, Fabio Rojas, and Clayton A. Davis. The social media response to Black Lives Matter: how Twitter users interact with Black Lives Matter through hashtag use. *Ethnic and Racial Studies*, 40(11):1814–1830, 2017.

[40] Jordana J. George and Dorothy E. Leidner. From clicktivism to hacktivism: Understanding digital activism. *Information and Organization*, 29(3):100249, 2019.

[41] Ryan J. Gallagher, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PLoS ONE*, 13(4):1–23, 2018.

[42] Thomas Marlow, Sean Miller, and J. Timmons Roberts. Bots and online climate discourses: Twitter discourse on president trump' s announcement of u.s. withdrawal from the paris agreement. *Climate Policy*, 21(6):765–777, 2021.

[43] Chang-Feng Chen, Wen Shi, Jing Yang, and Hao-Huan Fu. Social bots' role in climate change discussion on twitter: Measuring standpoints, topics, and interaction strategies. *Advances in Climate Change Research*, 12(6):913–923, 2021.

[44] Extinction Rebellion. *This Is Not a Drill: An Extinction Rebellion Handbook*. Penguin Books, 2019. Accessed: 2024-10-09.

[45] Kevin Michael DeLuca. Extinction rebellion, image events, social media and the eclipse of the earth. *Social Anthropology/Anthropologie sociale*, 29(1):216–218, 2021.

[46] László Christián, Ákos Erdős, and Gergő Háló. The background and repercussions of the george floyd case. *Cogent Social Sciences*, 8(1):2082094, 2022.

[47] University of Oxford. Research ethics: Best practice guidance, 2024. Accessed: 2024-10-06.

[48] Raya A Jones. What makes a robot 'social' ? *Social studies of science*, 47(4):556–579, 2017.

[49] Douglas Guilbeault. Growing bot security: An ecological view of bot agency. *International Journal of Communication*, 10(June):5003–5021, 2016.

[50] Andrej Duh, Marjan Slak Rupnik, and Dean Korošak. Collective Behavior of Social Bots Is Encoded in Their Temporal Twitter Activity. *Big Data*, 6(2):113–123, 2018.

[51] Carlene Lebeuf, Alexey Zagalsky, Matthieu Foucault, and Margaret-Anne Storey. Defining and classifying software bots: A faceted taxonomy. In *2019 IEEE/ACM 1st international workshop on bots in software engineering (BotSE)*, pages 1–6. IEEE, 2019.

[52] Linda Erlenhov, Francisco Gomes de Oliveira Neto, Riccardo Scandariato, and Philipp Leitner. Current and future bots in software development. In *2019 IEEE/ACM 1st International Workshop on Bots in Software Engineering (BotSE)*, pages 7–11. IEEE, 2019.

[53] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*, pages 373–383. Springer, 2020.

[54] Elizabeth Stinson and John C Mitchell. Characterizing bots' remote control behavior. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 89–108. Springer, 2007.

[55] Stefan Stieglitz, Florian Brachten, Björn Ross, and Anna-Katharina Jung. Do social bots dream of electric sheep? a categorisation of social media bot accounts. *arXiv preprint arXiv:1710.04044*, 2017.

[56] Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. Detection of bots in social media: a systematic review. *Information Processing & Management*, 57(4):102250, 2020.

[57] Carolina Alves De Lima Salge and Elena Karahanna. Protesting Corruption on Twitter: Is It a Bot or Is It a Person? *Academy of Management Discoveries*, 4(1):32–49, 2018.

[58] Magdalena Wischnewski, Thao Ngo, Rebecca Bernemann, Martin Jansen, and Nicole Krämer. "i agree with you, bot!" how users (dis) engage with social bots on twitter. *New Media & Society*, 26(3):1505–1526, 2024.

[59] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

[60] Stefano De Paoli. The raise of the robots in virtual worlds: A comparison and a framework for investigating bots in social networks sites and mmogs. In *Handbook on 3D3C platforms: applications and tools for three dimensional systems for community, creation and commerce*, pages 59–83. Springer, 2016.

[61] Milena Tsvetkova, Taha Yasseri, Eric T. Meyer, J. Brian Pickering, Vegard Engen, Paul Walland, Marika Lüders, Asbjørn Følstad, and George Bravos. Understanding Human-Machine Networks. *ACM Computing Surveys*, 50(1):1–35, 2017.

[62] Richard J. Oentaryo, Arinto Murdopo, Philips K. Prasetyo, and Ee Peng Lim. On profiling bots in social media. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10046 LNCS:92–109, 2016.

[63] Gregory Maus. A typology of socialbots (abbrev.). In *Proceedings of the 2017 ACM on Web Science Conference*, pages 399–400, 2017.

[64] Nick Monaco and Samuel Woolley. *Bots.* John Wiley & Sons, 2022.

[65] Massimo Stella, Marco Cristoforetti, and Manlio De Domenico. Influence of augmented humans in online interactions during voting events. *PLoS ONE*, 14(5):1–8, 2019.

[66] Sofia Hurtado, Poushali Ray, and Radu Marculescu. Bot detection in reddit political discussion. In *Proceedings of the fourth international workshop on social sensing*, pages 30–35, 2019.

[67] Robert Gorwa and Douglas Guilbeault. Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy and Internet*, 12(2):225–248, 2020.

[68] Bello Shehu Bello, Reiko Heckel, and Leandro Minku. Reverse Engineering the Behaviour of Twitter Bots. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 27–34. IEEE, 2018.

[69] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 280–289, 2017.

[70] Zixuan Weng and Aijun Lin. Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, 19(24), 2022.

[71] Luca Luceri, Adam Badawy, Ashok Deb, and Emilio Ferrara. Red bots do it better: Comparative analysis of social bot partisan behavior. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, pages 1007–1012, 2019.

[72] Franziska Martini, Paul Samula, Tobias R Keller, and Ulrike Klinger. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data & Society*, 8(2):205395172110335, 2021.

[73] Salvador Lopez-Joya, J Angel Diaz-Garcia, M Dolores Ruiz, and Maria J Martin-Bautista. Bot detection in twitter: An overview. In *International Conference on Flexible Query Answering Systems*, pages 131–144. Springer, 2023.

[74] Ludovic Rheault and Andreea Musulan. Efficient detection of online communities and social bot activity during electoral campaigns. *Journal of Information Technology and Politics*, 18(3):324–337, 2021.

[75] Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political communication*, 37(2):256–280, 2020.

[76] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269, 2022.

[77] Jing Wang and Ioannis Ch Paschalidis. Botnet Detection Based on Anomaly and Community Detection. *IEEE Transactions on Control of Network Systems*, 4(2):392–404, 2017.

[78] Jiawei Zhou, Zhiying Xu, Alexander M Rush, and Minlan Yu. Automating botnet detection with graph neural networks. *arXiv preprint arXiv:2003.06344*, 2020.

[79] Michele Cirillo, Mario Di Mauro, Vincenzo Matta, and Marco Tambasco. Botnet identification in ddos attacks with multiple emulation dictionaries. *IEEE Transactions on Information Forensics and Security*, 16:3554–3569, 2021.

[80] Norah Abokhodair, Daisy Yoo, and David W. McDonald. Dissecting a social Botnet: Growth, content and influence in twitter. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, pages 839–851, 2015.

[81] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. The Socialbot Network: When bots socialize for fame and money. *ACM International Conference Proceeding Series*, pages 93–102, 2011.

[82] Carlos Freitas, Fabrício Benevenuto, Adriano Veloso, and Saptarshi Ghosh. An empirical study of socialbot infiltration strategies in the twitter social network. *Social Network Analysis and Mining*, 6(1), 2016.

[83] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*, 2012.

[84] Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities. *Proceedings of the 12th International Conference on Management of Digital EcoSystems, MEDES 2020*, pages 55–63, 2020.

[85] Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 25–32. IEEE, 2010.

[86] Cliff Changchun Zou and Ryan Cunningham. Honeypot-aware advanced botnet construction and maintenance. In *International Conference on Dependable Systems and Networks (DSN'06)*, pages 199–208. IEEE, 2006.

[87] Ping Wang, Lei Wu, Ryan Cunningham, and Cliff C Zou. Honeypot detection in advanced botnet attacks. *International Journal of Information and Computer Security*, 4(1):30–51, 2010.

[88] Fatih Haltaş, Erkam Uzun, Necati Şişeci, Abdulkadir Poşul, and Bakır Emre. An automated bot detection system through honeypots for large-scale. In *2014 6th International Conference On Cyber Conflict (CyCon 2014)*, pages 255–270. IEEE, 2014.

[89] Berta García-Orosa, Pablo Gamallo, Patricia Martín-Rodilla, and Rodrigo Martínez-Castaño. Hybrid Intelligence Strategies for Identifying, Classifying and Analyzing Political Bots. *Social Sciences*, 10(10):357, 2021.

[90] Grażyna Suchacka, Alberto Cabri, Stefano Rovetta, and Francesco Masulli. Efficient on-the-fly web bot detection. *Knowledge-Based Systems*, 223:107074, 2021.

[91] Abbas Abou Daya, Mohammad A Salahuddin, Noura Limam, and Raouf Boutaba. Botchase: Graph-based bot detection using machine learning. *IEEE Transactions on Network and Service Management*, 17(1):15–29, 2020.

[92] A Jo. The promise and peril of generative ai. *Nature*, 614(1):214–216, 2023.

[93] Emilio Ferrara. Social bot detection in the age of chatgpt: Challenges and opportunities. *First Monday*, 28(6), 2023.

[94] Ross Schuchard, Andrew T. Crooks, Anthony Stefanidis, and Arie Croitoru. Bot stamina: examining the influence and staying power of bots in online social networks. *Applied Network Science*, 4(1), 2019.

[95] D Arnaudo. Computational propaganda in brazil: social bots during elections. Accessed: [09-10-2024], 2017.

[96] Rose Marie Santini, Débora Salles, and Giulia Tucci. When Machine Behavior Targets Future Voters: The Use of Social Bots to Test Narratives for Political Campaigns in Brazil. *International Journal of Communication*, 15:1220–1243, 2021.

[97] Fabian Schäfer, Stefan Evert, and Philipp Heinrich. Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinzō Abe's Hidden Nationalist Agenda. *Big Data*, 5(4):294–309, 2017.

[98] E. Kušen and M. Strembeck. Why so emotional? an analysis of emotional bot-generated content on twitter. In *Proceedings of the 3rd International Conference on Complexity, Future Information Systems and Risk*, volume 1, pages 13–22, March 2018.

[99] Henrich R Greve, Hayagreeva Rao, Paul Vicinanza, and Echo Yan Zhou. Online conspiracy groups: Micro-bloggers, bots, and coronavirus conspiracy talk on twitter. *American Sociological Review*, 87(6):919–949, 2022.

[100] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Political science: Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, 2019.

[101] Peter Suciu. Bots and misinformation on social media could undermine the midterm elections. https://www.forbes.com/sites/petersuciu/2022/11/02/bots-and-misinformation-on-social-media-could-undermine-the-midterm-elections/, 2022. Accessed: 2024-10-09.

[102] James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A Makse, and Boleslaw K Szymanski. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. *Nature Human Behaviour*, pages 1–13, 2023.

[103] Eric Tadeu Camacho De Oliveira, Fabricio Olivetti De Franca, Denise Hideko Goya, and Claudio Luis De Camargo Penteado. The influence of retweeting robots during brazilian protests. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2016-March:2068–2076, 2016.

[104] Bence Kollanyi. Where do bots come from? An analysis of bot codes shared on GitHub. *International Journal of Communication*, 10(June):4932–4951, 2016.

[105] Mairieli Wessel. Leveraging software bots to enhance developers' collaboration in online programming communities. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pages 183–188, 2020.

[106] Andrea Millimaggi and Florian Daniel. On twitter bots behaving badly: A manual and automated analysis of python code patterns on github. *Journal of Web Engineering*, 18(8):801–835, 2019.

[107] Lena Frischlich, Niels Göran Mede, and Thorsten Quandt. The markets of manipulation: The trading of social bots on clearnet and darknet markets. In *Disinformation in Open Online Media: First Multidisciplinary International Symposium, MISDOOM 2019, Hamburg, Germany, February 27–March 1, 2019, Revised Selected Papers 1*, pages 89–100. Springer, 2020.

[108] Johan Farkas and Marco Bastos. Ira propaganda on twitter: Stoking antagonism and tweeting local news. In *Proceedings of the 9th International Conference on social media and society*, pages 281–285, 2018.

[109] Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. Political effects of the internet and social media. *Annual Review of Economics*, 12:415–438, 2020.

[110] Taha Yasseri, Helen Margetts, Peter John, and Scott Hale. *Political turbulence: How social media shape collective action*. Princeton University Press, 2015.

[111] Fei Xiong, Yun Liu, Zhen-jiang Zhang, Jiang Zhu, and Ying Zhang. An information diffusion model based on retweeting mechanism for online social media. *Physics letters A*, 376(30-31):2103–2108, 2012.

[112] Jennifer Earl. The dynamics of protest-related diffusion on the web. *Information, communication & society*, 13(2):209–225, 2010.

[113] Sandra González-Bailón, Javier Borge-Holthoefer, and Yamir Moreno. Broadcasters and Hidden Influentials in Online Protest Diffusion. *American Behavioral Scientist*, 57(7):943–965, 2013.

[114] W. Russell Neuman, Lauren Guggenheim, S. Mo Jang, and Soo Young Bae. The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. *Journal of Communication*, 64(2):193–214, 2014.

[115] Andrea Ceron, Luigi Curini, and Stefano M. Iacus. First- and second-level agenda setting in the Twittersphere: An application to the Italian political debate. *Journal of Information Technology and Politics*, 13(2):159–174, 2016.

[116] Raymond A. Harder, Julie Sevenans, and Peter Van Aelst. Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times. *International Journal of Press/Politics*, 22(3):275–293, 2017.

[117] Grażyna Piechota. The transnational discourse of political protests: Setting the agenda through social media. *Kome*, 9(1):19–40, 2021.

[118] Jing Guo and Shujun Liu. From #blacklivesmatter to #stopasianhate: Examining network agenda-setting effects of hashtag activism on twitter. *Social Media+ Society*, 8(4):20563051221146182, 2022.

[119] Manju Ahuja, Pankaj Patel, and Ayoung Suh. The influence of social media on collective action in the context of digital activism: An affordance approach. In *Proceedings of the Hawaii International Conference on System Sciences*, 2024.

[120] G. Harindranath, Edward Bernroider, and Sherif Kamel. Social media and social transformation movements: The role of affordances and platforms. In *Proceedings*

*of the European Conference on Information Systems (ECIS) 2015*, Münster, Germany, 2015. Association for Information Systems.

[121] Yannis Theocharis, Shelley Boulianne, Karolina Koc-Michalska, and Bruce Bimber. Platform affordances and political participation: how social media reshape political engagement. *West European Politics*, 46(4):788–811, 2023.

[122] Sandra González-Bailón and Ning Wang. Networked discontent: The anatomy of protest campaigns in social media. *Social networks*, 44:95–104, 2016.

[123] Ion Bogdan Vasi and Chan S. Suh. Online Activities, Spatial Proximity, and the Diffusion of the Occupy Wall Street Movement in the United States. *Mobilization: An International Quarterly*, 21(2):139–154, 2016.

[124] Zachary C. Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):1–9, 2015.

[125] Sarah J Jackson and Brooke Foucault Welles. # ferguson is everywhere: Initiators in emerging counterpublic networks. *Information, Communication & Society*, 19(3):397–418, 2016.

[126] Eva Anduiza, Camilo Cristancho, and José M. Sabucedo. Mobilization through online social networks: the political protest of the indignados in Spain. *Information Communication and Society*, 17(6):750–764, 2014.

[127] Deen Freelon, Charlton D McIlwain, and Meredith Clark. Beyond the hashtags:# ferguson,# blacklivesmatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University, Forthcoming*, 2016.

[128] Emilio Ferrara. Measuring social spam and the effect of bots on information diffusion in social media. *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, pages 229–255, 2018.

[129] Björn Ross, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412, 2019.

[130] Kenneth T Andrews and Michael Biggs. The dynamics of protest diffusion: Movement organizations, social networks, and news media in the 1960 sit-ins. *American sociological review*, 71(5):752–777, 2006.

[131] Jeroen Van Laer. Activists online and offline: The Internet as an information channel for protest demonstrations. *Mobilization*, 15(3):347–366, 2010.

[132] Shelley Boulianne. Social media use and participation: a meta-analysis of current research. *Information Communication and Society*, 18(5):524–538, 2015.

[133] Zeynep Tufekci. *Twitter and tear gas: The power and fragility of networked protest.* Yale University Press, 2017.

[134] Brandon Brooks, Bernie Hogan, Nicole Ellison, Cliff Lampe, and Jessica Vitak. Assessing structural correlates to social capital in Facebook ego networks. *Social Networks*, 38(1):1–15, 2014.

[135] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin I.M. Dunbar. Online Social Networks and information diffusion: The role of ego networks. *Online Social Networks and Media*, 1:44–55, 2017.

[136] Jennifer M. Larson, Jonathan Nagler, Jonathan Ronen, and Joshua A. Tucker. Social Networks and Protest Participation: Evidence from 130 Million Twitter Users. *American Journal of Political Science*, 00(00):1–16, 2019.

[137] Rosemary Clark-Parsons. "i see you, i believe you, i stand with you" :# metoo and the performance of networked feminist visibility. *Feminist Media Studies*, 21(3):362–380, 2021.

[138] Marcia Mundt, Karen Ross, and Charla M Burnett. Scaling social movements through social media: The case of black lives matter. *Social media+ society*, 4(4), 2018.

[139] Sandra González-Bailón, Ning Wang, and Javier Borge-Holthoefer. The emergence of roles in large-scale networks of communication. *EPJ Data Science*, 3:1–16, 2014.

[140] Pablo Barberá, Ning Wang, Richard Bonneau, John T. Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. The Critical Periphery in the Growth of Social Protests. *PLOS ONE*, 10(11):e0143611, nov 2015.

[141] Carlos Lemos, Helder Coelho, and Rui J Lopes. Agent-based modeling of protests and violent confrontation: a micro-situational, multi-player, contextual rule-based approach. *Agent-Based modeling of protests and violent confrontation: A micro-situational, multi-player, contextual rule-based approach*, pages 136–161, 2014.

[142] Bruce Etling. *Network structure, brokerage, and framing: how the internet and social media facilitate high-risk collective action*. PhD thesis, University of Oxford, 2016.

[143] Michael Rossetti and Tauhid Zaman. Bots, disinformation, and the first impeachment of U.S. President Donald Trump. *PLoS ONE*, 18(5 MAY):1–27, 2023.

[144] D. Krackhardt, N. Nohria, and R. G. Eccles. The strength of strong ties. In C. E. M. and A. N., editors, *Networks in the Knowledge Economy*, page 82. Oxford University Press, Oxford, 2003.

[145] T. Y. T. Tang. Collective memories, emotions, and spatial tactics in social movements: The case of the umbrella movement in hong kong. *Emotion, Space and Society*, 38:100767, 2021.

[146] Eeva Luhtakallio and Mattias Wahlström. Protest for a future ii: Composition, mobilization and motives of the participants in fridays for future climate protests on 20-27 september, 2019, in 19 cities around the world. Working paper, Open Society Foundations (OSF), United Kingdom, 2020.

[147] Lisa Mueller. Crowd Cohesion and Protest Outcomes. *American Journal of Political Science*, 00(0):1–16, 2022.

[148] R. Abdulla. Social media, the public sphere, and public space in authoritarian settings: A case study of egypt' s tahrir square and 18 years of user generated content. *Social Media & Society*, 9(1), 2023.

[149] A. Kachen, A. S. Krishen, M. Petrescu, R. D. Gill, and P. C. Peter. #metoo, #methree, #mefour: Twitter as community building across academic and corporate institutions. *Psychology & Marketing*, 38(3):455–469, 2021.

[150] S. S. Duvall. Too famous to protest: Far-right online community bonding over collective desecration of colin kaepernick, fame, and celebrity activism. *Journal of Communication Inquiry*, 44(3):256–278, 2020.

[151] Robert D. Benford and David A. Snow. Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26(1):611–639, 2000.

[152] Henry Farrell. The Consequences of the Internet for Politics. *Annual Review of Political Science*, 15(1):35–52, 2012.

[153] M. Wahlström, M. Wennerhag, and C. Rootes. Framing "the climate issue" : Patterns of participation and prognostic frames among climate summit protesters. *Global Environmental Politics*, 13(4):101–122, 2013.

[154] A. Svensson and M. Wahlström. Climate change or what? prognostic framing by fridays for future protesters. *Social Movement Studies*, 22(1):1–22, 2023.

[155] Y. Xiong, M. Cho, and B. Boatwright. Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of twitter during the #metoo movement. *Public Relations Review*, 45(1):10–23, 2019.

[156] Ryan J Gallagher, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. *PloS one*, 13(4):e0195644, 2018.

[157] Ugochukwu Etudo, Victoria Yoon, and Niam Yaraghi. From facebook to the streets: Russian troll ads and black lives matter protests. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, Hawaii, USA, 2019. Hawaii International Conference on System Sciences.

[158] Kasey Stricklin and Megan K. McBride. Social media bots: Laws, regulations, and platform policies. Technical report, Center for Naval Analyses, 2020. Accessed: 2024-10-09.

[159] R. Stuart Geiger. Bots, bespoke, code and the materiality of software platforms. *Information Communication and Society*, 17(3):342–356, 2014.

[160] Aaron Delwiche. Computational Propaganda and the Rise of the Fake Audience. In *The SAGE Handbook of Propaganda*, pages 105–125. SAGE Publications, 2020.

[161] Sanket Barhate, Ritika Mangla, Dhwani Panjwani, Suyash Gatkal, and Faruk Kazi. Twitter bot detection and their influence in hashtag manipulation. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–7. IEEE, 2020.

[162] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[163] Simone Stolzoff. The problem with social media has never been about bots. it' s always been about business models. `https://qz.com/1449402/how-to-solve-social-medias-bot-problem/`, 2018. [Accessed 11-09-2024].

[164] X. X automation rules. `https://help.x.com/en/rules-and-policies/x-automation`, 2024. Accessed: 2024-09-11.

[165] Judith E Rosenbaum and Gwen Bouvier. Twitter, social movements and the logic of connective action: Activism in the 21st century–an introduction. *Participation: Journal of Audience & Reception Studies*, 17(1):120–125, 2020.

[166] Taina Bucher, Anne Helmond, et al. The affordances of social media platforms. *The SAGE handbook of social media*, 1:233–254, 2018.

[167] Steffi Siegert, Mikael Holmgren Caicedo, and Maria Mårtensson Hansson. Boundaryless twitter use: On the affordances of social media. *Social Sciences*, 9(11):201, 2020.

[168] Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. Brevity is the soul of Twitter: The constraint affordance and political discussion. *Journal of Communication*, 69(4):345–372, 2019.

[169] Chamil Rathnayake and Daniel D Suthers. Twitter issue response hashtags as affordances for momentary connectedness. In *Proceedings of the 8th International Conference on Social Media & Society*, pages 1–10, 2017.

[170] Emmanuelle Vaast, Hani Safadi, Liette Lapointe, and Bogdan Negoita. Social media affordances for connective action. *MIS quarterly*, 41(4):1179–1206, 2017.

[171] B. Hanckel, S. Vivienne, P. Byron, B. Robards, and B. Churchill. 'that' s not necessarily for them' : Lgbtiq+ young people, social media platform affordances and identity curation. *Media, Culture & Society*, 41(8):1261–1278, 2019.

[172] E. Khazraee and A. N. Novak. Digitally mediated protest: Social media affordances for collective identity construction. *Social Media + Society*, 4(1), 2018.

[173] Oscar Berglund and Daniel Schmidt. *Extinction rebellion and climate change activism: Breaking the law to change the world*. Springer, 2020.

[174] Karen Bell and Gnisha Bevan. Beyond inclusion? perceptions of the extent to which extinction rebellion speaks to, and for, black, asian and minority ethnic (bame) and working-class communities. *Local environment*, 26(10):1205–1220, 2021.

[175] Janette Huttunen. The extinction rebellion and going beyond politics: can deliberation be the answer in climate mitigation? *Fennia-International Journal of Geography*, 202(1):163–168, 2024.

[176] Extinction Rebellion. Global hunger strike, 2023. Accessed: 2024-10-09.

[177] Martha Kirby. Modelling the fall and rise in the importance of the environment to the British public: 2006–2019. *British Journal of Politics and International Relations*, 2022.

[178] Joost de Moor, Michiel De Vydt, Katrin Uba, and Mattias Wahlström. New kids on the block: taking stock of the recent cycle of climate activism. *Social Movement Studies*, 20(5):619–625, 2021.

[179] Silas Harrebye. Global civil society and international summits: New labels for different types of activism at the cop15. *Journal of Civil Society*, 7(4):407–426, 2011.

[180] Stefan Gaarsmand Jacobsen and Oliver Hunt. Climate justice and global politics. In *Handbook of Critical Environmental Politics*, pages 192–205. Edward Elgar Publishing, 2022.

[181] Anna Friberg. Disrupting the present and opening the future: Extinction rebellion, fridays for future, and the disruptive utopian method. *Utopian studies: The journal of the society for utopian studies*, 33(1):1–17, 2022.

[182] Joost De Moor. The 'efficacy dilemma' of transnational climate activism: the case of cop21. *Environmental politics*, 27(6):1079–1100, 2018.

[183] Aron Buzogány and Patrick Scherhaufer. Framing different energy futures? comparing fridays for future and extinction rebellion in germany. *Futures*, 137:102904, 2022.

[184] A. Friberg. On the need for (con)temporary utopias: Temporal reflections on the climate rhetoric of environmental youth movements. *Time & Society*, 31(1):48–68, 2022.

[185] Tom Smiles and Gareth AS Edwards. How does extinction rebellion engage with climate justice? a case study of xr norwich. *Local environment*, 26(12):1445–1460, 2021.

[186] Diana Stuart. Radical hope: Truth, virtue, and hope for what is left in extinction rebellion. *Journal of Agricultural and Environmental Ethics*, 33(3):487–504, 2020.

[187] Duncan Weaver. Profile: Extinction rebellion: greening vanguardism? *Social Movement Studies*, 23(1):113–121, 2024.

[188] Neil Gunningham. Averting climate catastrophe: environmental activism, extinction rebellion and coalitions of influence. *King's Law Journal*, 30(2):194–202, 2019.

[189] Luke D Bevan. Climate change strategic narratives in the united kingdom: emergency, extinction, effectiveness. *Energy research & social science*, 69:101580, 2020.

[190] Garrett Chase. The early history of the black lives matter movement, and the implications thereof. *Nev. LJ*, 18:1091, 2017.

[191] Black Lives Matter. About, 2023. [Accessed 09-10-2024].

[192] Laurie Collier Hillstrom. *Black Lives Matter: From a moment to a movement.* Bloomsbury Publishing USA, 2018.

[193] Ho-Chun Herbert Chang, Allissa Richardson, and Emilio Ferrara. # justiceforgeorgefloyd: how instagram facilitated the 2020 black lives matter protests. *PLoS one*, 17(12):e0277864, 2022.

[194] Zackary Okun Dunivin, Harry Yaojun Yan, Jelani Ince, and Fabio Rojas. Black lives matter protests shift public discourse. *Proceedings of the National Academy of Sciences*, 119(10):e2117320119, 2022.

[195] Kevin Drakulich and Megan Denver. The partisans and the persuadables: Public views of black lives matter and the 2020 protests. *Perspectives on Politics*, 20(4):1191–1208, 2022.

[196] Amanda Nell Edgar and Andre E Johnson. *The struggle over black lives matter and all lives matter*. Rowman & Littlefield, 2020.

[197] The Guardian. George floyd protests latest news: rioters set fires as trump threatens to use military, 2020. [Accessed 09-09-2024].

[198] Janani Umamaheswar. Policing and racial (in) justice in the media: Newspaper portrayals of the "black lives matter" movement. *Civic Sociology*, 1(1), 2020.

[199] Kevin Drakulich, Kevin H Wozniak, John Hagan, and Devon Johnson. Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics. *Criminology*, 58(2):370–402, 2020.

[200] Mathis Ebbinghaus, Nathan Bailey, and Jacob Rubel. The effect of the 2020 black lives matter protests on police budgets: How "defund the police" sparked political backlash. *Social Problems*, page spae004, 2024.

[201] Megan Ming Francis and Leah Wright-Rigueur. Black lives matter in historical perspective. *Annual Review of Law and Social Science*, 17(1):441–458, 2021.

[202] Christopher J Lebron. *The making of black lives matter: A brief history of an idea*. Oxford University Press, 2023.

[203] Alvin B Tillery. What kind of movement is black lives matter? the view from twitter. *Journal of Race, Ethnicity, and Politics*, 4(2):297–323, 2019.

[204] Yuan Hsiao. Evaluating the mobilization effect of online political network structures: A comparison between the black lives matter network and ideal type network configurations. *Social Forces*, 99(4):1547–1574, 2021.

[205] Maegan Parker Brooks. Network matters: Black lives and blue lives advocacy in on and offline settings. In *Networking Argument*, pages 61–66. Routledge, 2019.

[206] W. Lance Bennett, Alexandra Segerberg, and Yunkang Yang. The strength of peripheral networks: Negotiating attention and meaning in complex media ecologies. *Journal of Communication*, 68(4):830, 2018.

[207] L. G. Stewart, A. Arif, A. C. Nied, E. S. Spiro, and K. Starbird. Drawing the lines of contention: Networked frame contests within #blacklivesmatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–23, 2017.

[208] Sander van Haperen, Justus Uitermark, and Walter Nicholls. The swarm versus the grassroots: places and networks of supporters and opponents of black lives matter on twitter. *Social Movement Studies*, 22(2):171–189, 2023.

[209] Spring-Serenity Duvall and Nicole Heckemeyer. # blacklivesmatter: black celebrity hashtag activism and the discursive formation of a social movement. *Celebrity studies*, 9(3):391–408, 2018.

[210] Summer Harlow and Anna Benbrook. How# blacklivesmatter: exploring the role of hip-hop celebrities in constructing racial identity on black twitter. *Information, communication & society*, 22(3):352–368, 2019.

[211] Keenan Jones, Jason RC Nurse, and Shujun Li. Out of the shadows: analyzing anonymous' twitter resurgence during the 2020 black lives matter protests. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 417–428, 2022.

[212] Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. Blacklivesmatter 2020: an analysis of deleted and suspended users in twitter. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 290–295, 2022.

[213] Twitter. Twitter api: Academic research. `https://developer.twitter.com/en/products/twitter-api/academic-research`, 2023. [Accessed 09-10-2024].

[214] The Verge. Twitter' s new Elon Musk API policy is already chilling academic research. `https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research`, May 2023.

[215] Kai Cheng Yang, Onur Varol, Clayton A. Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

[216] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[217] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[218] OpenAI. ChatGPT: A large language model trained by openai. https://openai.com/chatgpt, 2024. Accessed on March 13, 2024.

[219] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*, 2023.

[220] Lifang Wu, Yan Chen, Jie Yang, Ge Shi, Xinmu Qi, and Sinuo Deng. Event-centric opinion mining via in-context learning with chatgpt. In *China Conference on Knowledge Graph and Semantic Computing*, pages 83–94. Springer, 2023.

[221] Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, 2023.

[222] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[223] Tong Zhan, Chenxi Shi, Yadong Shi, Huixiang Li, and Yiyu Lin. Optimization techniques for sentiment analysis based on llm (gpt-3). *arXiv preprint arXiv:2405.09770*, 2024.

[224] Ahmed Alsayat. Improving sentiment analysis for social media applications using an ensemble deep learning language model. *Arabian Journal for Science and Engineering*, 47(2):2499–2511, 2022.

[225] Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. Towards interpreting topic models with chatgpt. In *The 20th World Congress of the International Fuzzy Systems Association*, 2023.

[226] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[227] Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14:100157, 2022.

[228] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. a Bitem Topic Model for Short Texts. *Www*, pages 1445–1455, 2013.

[229] Lei Shi, Gang Cheng, Shang Ru Xie, and Gang Xie. A word embedding topic model for topic detection and summary in social networks. *Measurement and Control (United Kingdom)*, 52(9-10):1289–1298, 2019.

[230] Jan Wijffels. *BTM: Biterm Topic Models for Short Text*, 2023. R package version 0.3.7.

[231] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[232] Maksim Terpilovskii and Christof Kälin. Bitermplus package, 2022.

[233] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery.

[234] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[235] Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, pages 165–174. Ieee, 2017.

[236] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20, 2019.

[237] Edoardo M Airoldi and Jonathan M Bischof. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, 111(516):1381–1403, 2016.

[238] Jonathan Bischof and Edoardo M Airoldi. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th international conference on machine learning (icml-12)*, pages 201–208, 2012.

[239] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42, 2017.

[240] Alexandra Schofield, Siqi Wu, Theo Bayard de Volo, Tatsuki Kuze, Alfredo Gomez, and Sharifa Sultana. " my very subjective human interpretation": Domain expert perspectives on navigating the text analysis loop for topic models. *Proceedings of the ACM on Human-Computer Interaction*, 9(1):1–30, 2025.

[241] Nils Köbis, Jean François Bonnefon, and Iyad Rahwan. Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6):679–685, 2021.

[242] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David G. Rand. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 118(7), feb 2021.

[243] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, sep 2018.

[244] Ian Lundberg, Rebecca Johnson, and Brandon M Stewart. What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565, 2021.

[245] Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

[246] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.

[247] Javier Pastor-Galindo, Félix Gómez Mármol, and Gregorio Martínez Pérez. Profiling users and bots in Twitter through social media analysis. *Information Sciences*, 613:161–183, 2022.

[248] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The Social Roles of Bots. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–29, 2018.

[249] Samuel C. Woolley and Philip N. Howard. Political communication, Computational Propaganda, and autonomous agents: Introduction. *International Journal of Communication*, 10:4882–4890, 2016.

[250] Robert W. Gehl and Maria Bakardjieva. *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. Routledge, USA, 1st edition, 2016.

[251] Andrew Chadwick. *The hybrid media system: Politics and power*. Oxford University Press, 2017.

[252] Denis Stukal, Sergey Sanovich, Joshua A. Tucker, and Richard Bonneau. For Whom the Bot Tolls: A Neural Networks Approach to Measuring Political Orientation of Twitter Bots in Russia. *SAGE Open*, 9(2), 2019.

[253] Dhiraj Murthy, Alison B. Powell, Ramine Tinati, Nick Anstead, Leslie Carr, Susan J. Halford, and Mark Weal. Bots and political influence: A sociotechnical investigation of social network capital. *International Journal of Communication*, 10(June):4952–4971, 2016.

[254] Jennifer Earl, Thomas V Maher, and Jennifer Pan. The digital repression of social movements, protest, and activism: A synthetic review. *Science Advances*, 8(10):eabl8198, 2022.

[255] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, and Jon Crowcroft. A large-scale behavioural analysis of bots and humans on twitter. *ACM Transactions on the Web (TWEB)*, 13(1):1–23, 2019.

[256] Masumi Shirakawa, Takahiro Hara, and Takuya Maekawa. Never abandon minorities: Exhaustive extraction of bursty phrases on microblogs using set cover problem. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2358–2367, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[257] Carmela Comito, Agostino Forestiero, and Clara Pizzuti. Bursty event detection in twitter streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(4):1–28, 2019.

[258] Neil Carter. *The Politics of the Environment: Ideas, Activism, Policy*. Cambridge University Press, 3 edition, 2018.

[259] Claudia Wagner, Silvia Mitter, Christian Körner, Markus Strohmaier, et al. When social bots attack: Modeling susceptibility of users in online social networks. In *# MSM*, pages 41–48, 2012.

[260] Eiman Alothali, Nazar Zaki, Elfadil A Mohamed, and Hany Alashwal. Detecting social bots on twitter: a literature review. In *2018 International conference on innovations in information technology (IIT)*, pages 175–180. IEEE, 2018.

[261] K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002, 2008.

[262] By David Easley. Chapter 16 Information Cascades. *Networks, Crowds, and Markets*, pages 483–508, 2010.

[263] M Hashem Pesaran, Yongcheol Shin, and Richard J Smith. Structural analysis of vector error correction models with exogenous i (1) variables. *Journal of econometrics*, 97(2):293–343, 2000.

[264] Balazs Vedres and Orsolya Vasarhelyi. Gendered behavior as a disadvantage in open source software development. *EPJ Data Science*, 8(1):25, 2019.

[265] Wouter Van Atteveldt, Mariken ACG Van der Velden, and Mark Boukes. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140, 2021.

[266] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9), 2024.

[267] Kai-Cheng Yang and Filippo Menczer. Anatomy of an ai-powered malicious social botnet. *arXiv preprint arXiv:2307.16336*, 2023.

[268] The Verge. Reddit' s API changes are already starting to shut down some apps. `https://www.theverge.com/2023/6/9/23755640/reddit-api-changes-apps-apollo-shut-down-ama-spez-steve-huffman`, June 2023.

[269] TechCrunch. IRL shuts down, citing issues with fake users. `https://techcrunch.com/2023/06/26/irl-shut-down-fake-users/`, June 2023.

[270] Harry Yaojun Yan, Kai-Cheng Yang, James Shanahan, and Filippo Menczer. Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports*, 13(1):20707, 2023.

[271] European Commission. Digital Services Act (DSA) overview. `https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en#:~:text=Digital%20Services%20Act%20(DSA)%20overview,online%20travel%20and%20accommodation%20platforms.`, August 2023.

[272] Irene Solaiman. The gradient of generative ai release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 111–122, New York, NY, USA, 2023. Association for Computing Machinery.

[273] Siyu Li, Jin Yang, and Kui Zhao. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *arXiv preprint arXiv:2307.10337*, 2023.

[274] Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. Are large language models a threat to digital public goods? evidence from activity on stack overflow. *arXiv preprint arXiv:2307.07367*, 2023.

[275] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. *CoRR*, abs/2006.06867, 2020.

[276] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1096–1103, apr 2020.

[277] Mohammad Taha Bahadori and Yan Liu. An examination of practical granger causality inference. *Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013*, pages 467–475, 2013.

[278] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice H. Oh, editors, *ICWSM*. The AAAI Press, 2014.

[279] Panpan Xu, Yingcai Wu, Enxun Wei, Tai Quan Peng, Shixia Liu, Jonathan J H Zhu, and Huamin Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2012–2021, 2013.

[280] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3):554–559, 2016.

[281] Hamed Abdelhaq, Michael Gertz, and Christian Sengstock. Spatio-temporal characteristics of bursty words in twitter streams. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 194–203, 2013.

[282] Shelley Boulianne, Mireille Lalancette, and David Ilkiw. "school strike 4 climate" : Social media and the international youth protest on climate change. *Media and Communication*, 8(2):208–218, 2020.

[283] David E. Campbell. Social networks and political participation. *Annual Review of Political Science*, 16:33–48, 2013.

[284] Se Jung Park, Yon Soo Lim, and Han Woo Park. Comparing twitter and youtube networks in information diffusion: The case of the "occupy wall street" movement. *Technological forecasting and social change*, 95:208–217, 2015.

[285] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1(1):1–7, 2011.

[286] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *Plos one*, 13(4):e0196087, 2018.

[287] Dilli Bikram Edingo. Social media, public sphere and counterpublics: An exploratory analysis of the networked use of twitter during the protests against the citizenship amendment act in india. *The Journal of Social Media in Society*, 10(2):76–101, 2021.

[288] John Tawa, Ruqian Ma, and Shinji Katsumoto. "all lives matter" : The cost of colorblind racial attitudes in diverse social networks. *Race and Social Problems*, 8:196–208, 2016.

[289] Nicolas Guenon des Mesnards, David Scott Hunter, Zakaria El Hjouji, and Tauhid Zaman. Detecting Bots and Assessing Their Impact in Social Networks. *Operations Research*, 70(1):1–22, 2022.

[290] Yannis Theocharis and Jan W Van Deth. The continuous expansion of citizen participation: A new taxonomy. *European political science review*, 10(1):139–163, 2018.

[291] Pablo Suárez-Serrato, Margaret E Roberts, Clayton Davis, and Filippo Menczer. On the influence of social bots in online protests: preliminary findings of a mexican case study. In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II 8*, pages 269–278. Springer, 2016.

[292] Ashkan Dehghan, Kinga Siuta, Agata Skorupka, Akshat Dubey, Andrei Betlen, David Miller, Wei Xu, Bogumił Kamiński, and Paweł Prałat. Detecting bots in social-networks using node and structural embeddings. *Journal of Big Data*, 10(1):119, 2023.

[293] Linda Li, Orsolya Vásárhelyi, and Balázs Vedres. Social bots spoil activist sentiment without eroding engagement. *Scientific Reports*, 14(1):27005, 2024.

[294] Randall Wald, Taghi M. Khoshgoftaar, Amri Napolitano, and Chris Sumner. Predicting susceptibility to social bots on Twitter. *Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration, IEEE IRI 2013*, pages 6–13, 2013.

[295] Scott Appling and Erica J. Briscoe. The perception of social bots by human and machine. *FLAIRS 2017 - Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference*, pages 20–25, 2017.

[296] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. Acting the part: Examining information operations within# blacklivesmatter discourse. *Proceedings of the ACM on Human-computer Interaction*, 2(CSCW):1–27, 2018.

[297] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[298] Edi Surya Negara, Dendi Triadi, and Ria Andryani. Topic modelling twitter data with latent dirichlet allocation method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 386–390. IEEE, 2019.

[299] Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034, 2020.

[300] Noah E. Friedkin. Social cohesion. *Annual Review of Sociology*, 30:409–425, 2004.

[301] Sandra González-Bailón and Yphtach Lelkes. Do social media undermine social cohesion? A critical review. *Social Issues and Policy Review*, 17(1):155–180, 2023.

[302] Alejandro Portes and Erik Vickstrom. Diversity, social capital, and cohesion. *Annual Review of Sociology*, 37:461–479, 2011.

[303] James Moody and Douglas R White. Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. *American Sociological Review*, 68(1):103–127, 2003.

[304] J. Martí, M. Bolíbar, and C. Lozares. Network cohesion and social support. *Social Networks*, 48:192–201, 2017.

[305] David B Tindall. Social movement participation over time: An ego-network approach to micro-mobilization. *Sociological Focus*, 37(2):163–184, 2004.

[306] David Schiefer and Jolanda van der Noll. The Essentials of Social Cohesion: A Literature Review. *Social Indicators Research*, 132(2):579–603, 2017.

[307] Roslyn Dakin, Paisley Clunis, Emily M Cornthwaite, and T Brandt Ryder. Reciprocal social ties drive stability within a social network. *bioRxiv*, 2020. Preprint.

[308] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. Comparing methods for comparing networks. *Scientific Reports*, 9(1):1–19, 2019.

[309] Fatma Şen, R. Thomas Wigand, Nitin Agarwal, Ercan Tokdemir, and Robert Kasprzyk. Focal structures analysis: Identifying influential sets of individuals in a social network. *Social Network Analysis and Mining*, 6:1–22, 2016.

[310] Sebastián Valenzuela, Teresa Correa, and Homero G. de Zúñiga. Ties, likes, and tweets: Using strong and weak ties to explain differences in protest participation across facebook and twitter use. In *Studying Politics Across Media*, pages 117–134. Routledge, 2020.

[311] Lorien Jasny and Dana R Fisher. How networks of social movement issues motivate climate resistance. *Social Networks*, 75(October 2021):159–169, 2023.

[312] X. Yan, L. G. Jeub, A. Flammini, F. Radicchi, and S. Fortunato. Weight thresholding on complex networks. *Physical Review E*, 98(4):042304, 2018.

[313] C. Peng, T. G. Kolda, and A. Pinar. Accelerating community detection by using k-core subgraphs. *arXiv preprint arXiv:1403.2226*, 2014.

[314] A. Hagberg, P. J. Swart, and D. A. Schult. Exploring network structure, dynamics, and function using networkx. Technical Report LA-UR-08-05495, Los Alamos National Laboratory (LANL), Los Alamos, NM, United States, 2008.

[315] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Generalized louvain method for community detection in large networks. In *2011 11th international conference on intelligent systems design and applications*, pages 88–93. IEEE, 2011.

[316] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[317] A. P. Kartun-Giles and G. Bianconi. Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. *Chaos, Solitons & Fractals: X*, 1:100004, 2019.

[318] D. J. Kim, M. Salvacion, M. Salehan, and D. W. Kim. An empirical study of community cohesiveness, community attachment, and their roles in virtual community participation. *European Journal of Information Systems*, 32(3):573–600, 2023.

[319] B. Vedres and O. Vásárhelyi. Inclusion unlocks the creative potential of gender diversity in teams. *Scientific Reports*, 13(1):13757, 2023.

[320] Gary Alan Fine. Public narration and group culture: Discerning discourse in social movements. In *Social movements and culture*, pages 127–143. Routledge, 2013.

[321] Britta Baumgarten and Peter Ullrich. *Discourse, power, and governmentality. Social movement research with and beyond Foucault*. Springer, 2016.

[322] Samuel Merrill and Simon Lindgren. The rhythms of social movement memories: The mobilization of silvio meier' s activist remembrance across platforms. *Social Movement Studies*, 19(5-6):657–674, 2020.

[323] Yannis Theocharis, Will Lowe, Jan W Van Deth, and Gema García-Albacete. Using twitter to mobilize protest action: online mobilization patterns and action repertoires in the occupy wall street, indignados, and aganaktismenoi movements. *Information, Communication & Society*, 18(2):202–220, 2015.

[324] W Lance Bennett and Alexandra Segerberg. *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge University Press, 2013.

[325] Holly J McCammon, Courtney Sanders Muse, Harmony D Newman, and Teresa M Terrell. Movement framing and discursive opportunity structures: The political successes of the us women's jury movements. *American Sociological Review*, 72(5):725–749, 2007.

[326] Ruud Koopmans and Susan Olzak. Discursive opportunities and the evolution of right-wing violence in germany. *American journal of Sociology*, 110(1):198–230, 2004.

[327] Hamideh Molaei. Discursive opportunity structure and the contribution of social media to the success of social movements in indonesia. *Information, Communication & Society*, 18(1):94–108, 2015.

[328] Marcelo Mendoza, Maurizio Tesconi, and Stefano Cresci. Bots in social and interaction networks: detection and impact estimation. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–32, 2020.

[329] Olga Filatova and Daniil Volkovskii. The online discourse as a form of e-participation: the experience of internet discourse research. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, pages 326–333, 2020.

[330] Cristina Moreno-Almeida. Memes as snapshots of participation: The role of digital amateur activists in authoritarian regimes. *New Media & Society*, 23(6):1545–1566, 2021.

[331] Mattias Wahlström and Anton Törnberg. Social media mechanisms for right-wing political violence in the 21st century: Discursive opportunities, group dynamics, and co-ordination. *Terrorism and Political Violence*, 33(4):766–787, 2021.

[332] Sarah Glozer, Robert Caruana, and Sally A Hibbert. The never-ending story: Discursive legitimation in social media dialogue. *Organization Studies*, 40(5):625–650, 2019.

[333] Guobin Yang. Narrative agency in hashtag activism: The case of# blacklivesmatter. *Media and communication*, 4(4):13–17, 2016.

[334] A. F. Simon and J. Jerit. Toward a theory relating political discourse, media, and public opinion. *Journal of Communication*, 57(2):254–271, 2007.

[335] S. Almanssori and M. Stanley. Public pedagogy on sexual violence: A feminist discourse analysis of youtube vlogs after #metoo. *Journal of Curriculum and Pedagogy*, 19(3):254–277, 2022.

[336] Helen Margetts and Peter John. How rediscovering nodality can improve democratic governance in a digital world. *Public Administration*, 2023.

[337] Bob Van De Velde, Albert Meijer, and Vincent Homburg. Police message diffusion on twitter: analysing the reach of social media communications. *Behaviour & information technology*, 34(1):4–16, 2015.

[338] Jingwen Bian, Yang Yang, and Tat-Seng Chua. Predicting trending messages and diffusion participants in microblogging network. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 537–546, 2014.

[339] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the international AAAI conference on web and social media*, volume 4, 2010.

[340] Leonardo Castro-Gonzalez, Sukankana Chakraborty, Helen Margetts, Hardik Rajpal, Daniele Guariso, and Jonathan Bright. Who is driving the conversation? analysing the nodality of british mps and journalists on twitter. *arXiv preprint arXiv:2402.08765*, 2024.

[341] Katarzyna Gruszka and Madeleine Böhm. Out of sight, out of mind?(in) visibility of/in platform-mediated work. *new media & society*, 24(8):1852–1871, 2022.

[342] Kelley Cotter. Playing the visibility game: How digital influencers and algorithms negotiate influence on instagram. *New media & society*, 21(4):895–913, 2019.

[343] John A Noakes and Hank Johnston. Frames of protest: A road map to a perspective. *Frames of protest: Social movements and the framing perspective*, 1(11), 2005.

[344] Pauline Ketelaars. What strikes the responsive chord? the effects of framing qualities on frame resonance among protest participants. *Mobilization: An International Quarterly*, 21(3):341–360, 2016.

[345] James M. Jasper. Emotions and social movements: Twenty years of theory and research. *Annual Review of Sociology*, 37:285–303, 2011.

[346] Vincent Miller. Resonance as a social phenomenon. *Sociological Research Online*, 20(2):58–70, 2015.

[347] Belinda Robnett. Emotional resonance, social location, and strategic framing. *Sociological Focus*, 37(3):195–212, 2004.

[348] William J Coburn. Subjectivity, emotional resonance, and the sense of the real. *Psychoanalytic Psychology*, 18(2):303, 2001.

[349] Menghan Zhang, Ze Chen, Xue Qi, and Jun Liu. Could social bots' sentiment engagement shape humans' sentiment on covid-19 vaccine discussion on twitter? *Sustainability*, 14(9):5566, 2022.

[350] Wen Shi, Diyi Liu, Jing Yang, Jing Zhang, Sanmei Wen, and Jing Su. Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter. *International Journal of Environmental Research and Public Health*, 17(22), 2020.

[351] Thomas Poell. Social media, temporality, and the legitimacy of protest. *Social Movement Studies*, 19(5-6):609–624, 2020.

[352] Stefano Passini and Davide Morselli. Supporting protest movements: The effect of the legitimacy of the claims. *International Journal of Psychological Research*, 8(2):10–22, 2015.

[353] Kenneth T Andrews, Kraig Beyerlein, and Tuneka Tucker Farnum. The legitimacy of protest: explaining white southerners' attitudes toward the civil rights movement. *Social Forces*, 94(3):1021–1044, 2016.

[354] Christina Neumayer and Luca Rossi. Images of protest in social media: Struggle over visibility and visual narratives. *New Media & Society*, 20(11):4293–4310, 2018.

[355] Lina Dencik. Social media and the 'new authenticity' of protest. *Critical Perspectives on Social Media and Protest: Between Control and Emancipation. London: Rowman & Littlefield*, pages 203–218, 2015.

[356] Julie Ayling. A Contest for Legitimacy: The Divestment Movement and the Fossil Fuel Industry. *Law and Policy*, 39(4):349–371, 2017.

[357] E. F. Thomas and W. R. Louis. When will collective action be effective? violent and non-violent protests differentially influence perceptions of legitimacy and efficacy among sympathizers. *Personality and Social Psychology Bulletin*, 40(2):263–276, 2014.

[358] Jessica Einspänner, Mark Dang-Anh, and Caja Thimm. Computer-assisted content analysis of twitter data. In Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, editors, *Twitter and Society*, volume 89 of *Digital Formations*, pages 97–108. P. Lang, New York, 2014.

[359] Christian Baden, Neta Kligler-Vilenchik, and Moran Yarchi. Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, 14(3):165–183, 2020.

[360] Rémi Rampin and Vicky Rampin. Taguette: open-source qualitative data analysis. *Journal of Open Source Software*, 6(68):3522, 2021.

[361] Britta Grimme, Janina Pohl, Hendrik Winkelmann, Lucas Stampe, and Christian Grimme. Lost in transformation: Rediscovering llm-generated campaigns in social media. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pages 72–87. Springer, 2023.

[362] Helen Margetts. Rethinking democracy with social media. *Political Quarterly*, 90(S1), 2018.

[363] W. Lance Bennett and Barbara Pfetsch. Rethinking Political Communication in a Time of Disrupted Public Spheres. *Journal of Communication*, 68(2):243–253, 2018.

[364] Christina Ruess, Christian Pieter Hoffmann, Shelley Boulianne, and Katharina Heger. Online political participation: the evolution of a concept. *Information, Communication & Society*, 26(8):1495–1512, 2023.

[365] Giovanni C Santia, Munif Ishad Mujib, and Jake Ryland Williams. Detecting social bots on facebook in an information veracity context. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 463–472, 2019.

[366] Maxim Kolomeets and Andrey Chechulin. Analysis of the malicious bots market. In *2021 29th conference of open innovations association (FRUCT)*, pages 199–205. IEEE, 2021.

[367] Wei Fang and Chen Nie. Social media use, social bot literacy, perceived threats from bots, and perceived bot control: a moderated-mediation model. *Behaviour and Information Technology*, pages 1–17, 2023.

[368] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4485–4494, 2021.

[369] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[370] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[371] Emily Holden. Harvard and yale students disrupt football game for fossil fuel protest | us universities. `https://www.theguardian.com/us-news/2019/nov/23/harvard-yale-football-game-protest-fossil-fuels`, 2019. [Accessed 16-02-2023].

[372] BBC News. COP25: Thousands gather for change climate protests in Madrid — bbc.co.uk. https://www.bbc.co.uk/news/world-europe-50694361. [Accessed 07-09-2023].

[373] Damien Gayle. Extinction Rebellion bee protester glues himself to Lib Dem bus — theguardian.com. https://www.theguardian.com/environment/2019/dec/04/extinction-rebellion-protester-glues-himself-lib-dem-battlebus. [Accessed 07-09-2023].

[374] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[375] Laura K Nelson. Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1):3–42, 2020.

[376] Hjalmar Bang Carlsen and Snorre Ralund. Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society*, 9(1), 2022.

[377] Candice LaShara Edrington and Nicole Lee. Tweeting a social movement: Black lives matter and its use of twitter to share information, build community, and promote action. *The Journal of Public Interest Communications*, 2(2):289–289, 2018.

[378] Extinction Rebellion. Election rebellion: Twelve days of crisis. https://extinctionrebellion.uk/2019/11/26/election-rebellion-twelve-days-of-crisis/, Dec 2019.

[379] Tyler T Reny and Benjamin J Newman. The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 george floyd protests. *American political science review*, 115(4):1499–1507, 2021.

[380] Twitter. Api reference index | docs | twitter developer platform. https://developer.twitter.com/en/docs/api-reference-index, 2022. [Accessed 09-10-2024].

[381] Andrei P Kirilenko and Svetlana O Stepchenkova. Climate change discourse in mass media: application of computer-assisted content analysis. *Journal of Environmental Studies and Sciences*, 2:178–191, 2012.

[382] Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*, 2023.

[383] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), 2023.

[384] Dimitar Trajanov, Georgi Lazarev, Lubomir Chitkushev, and Irena Vodenska. Comparing the performance of chatgpt and state-of-the-art climate nlp models on climate-related text classification tasks. In *Proceedings of the 4th International Conference on Environmental Design (ICED2023)*, 2023.