


# Exploring Online Polarisation in the United States: A Computational Social Science Approach

Thesis submitted in partial fulfilment of the requirements for the degree of DPhil in Social Data Science at the Oxford Internet Institute, University of Oxford.

Word count: 35,400

Giuliano Formisano
Nuffield College, Oxford

Hilary Term 2025

**Abstract**

Affective polarisation—the profound divide between political groups—has emerged as a critical threat to American democracy, vividly illustrated by events like the January 6th 2021 Capitol Hill attack. Despite widespread concern, existing research has struggled to adequately measure polarisation at scale, trace its evolution across actors and time, and map its structural underpinnings in online networks. This thesis addresses these gaps by means of three papers, leveraging theoretical and methodological innovations, and large-scale empirical analysis of US Twitter data from 2015 to 2020.

The first paper introduces a novel framework for detecting polarising language using supervised machine learning classifiers based on transformer models. By integrating the political affiliations of human coders and benchmarking against generative artificial intelligence (AI), the study demonstrates that coder ideology shapes perceptions of divisiveness, with combined human coders' perspectives yielding more balanced and accurate classification than single-affiliation or AI-only models. These classifiers, which will be made publicly available, establish a robust foundation for scalable and nuanced measurement of affective polarisation.

Building on this measurement advance, the second paper investigates the temporal and directional dynamics of polarisation among US politicians, news media, and the public. Applying vector-autoregressive models to a comprehensive tweet dataset, the analysis reveals that polarising rhetoric is driven by a dynamic feedback loop: politicians, media, and citizens alternately lead and respond to shifts in divisive language. The media acts as a catalyst, while cross-party interactions often escalate polarisation, underscoring the complex, reciprocal nature of digital political discourse.

The third paper turns to the network structures that sustain elite polarisation. Using hierarchical exponential random graph models, it maps the relational architecture of political communication. It finds that party-based homophily dominates network formation. Democrat politicians hold stronger connections to other powerful nodes in the network than Republicans. The most divisive language is concentrated in the rare ties bridging partisan clusters, highlighting the joint role of network structure and rhetoric in entrenching polarisation.

Together, these studies advance both the measurement and substantive understanding of affective polarisation in digital public spheres, offering new tools and theoretical insights for researchers, policymakers, and platform designers seeking to address the challenges of polarisation in contemporary democracies.

# Acknowledgements

While only a Sith deals in absolutes, this thesis may be filled with hedging, several "it depends," and three chapters dedicated to explaining why the data might not be wrong. Yet, despite all this caution, I owe an immense debt to some amazing people who have contributed to every step of this journey. Their guidance, encouragement, and support have made this work possible.

Foremost, I would like to express my deepest gratitude to my supervisors, Professors Ralph Schroeder and Spyros Kosmidis, for their guidance and support throughout my academic journey. Ralph's expertise in internet studies and his thoughtful feedback during our many discussions have been invaluable to my research. Spyros, who has mentored me since my master's, has profoundly shaped my understanding and passion for political science. Coming from a small village in Cornwall to Oxford, I am especially grateful for the knowledge, encouragement, and enthusiasm he has shared with me. I am truly indebted to both for inspiring my love for this field and for their unwavering commitment to my development as a scholar.

My doctoral degree could not have been possible without the funding of the Economic and Social Research Council and Nuffield College, Oxford.

I thank my mentors and co-authors over the last few years for teaching me so much: Professors/Drs Tarik Abou-Chadi, José Manuel Aburto, Constantine Boussalis, Paul Chaisty, Pietro Castelli-Gattinara, Zenobia Chang, Anna Clemente, Elias Dinas, Luciano Floridi, Caterina Froio, Joerg Friedrichs, Ezequiel Gonzalez-Ocantos, Jane Green, Desmond King, Joshua Kalla, Helen Margetts, Vicky Nash, Clare Saunders, Florian Schaffner, Sascha Riaz, Lamprini Rori, David Rueda, Nelson Ruiz, Jas Sekhon, Miriam Sorace, Fabian Stephany, Niklas Stoehr, Yannis Theocaris, Ian Turner, Vicente Valentim, and Stephen Whitefield.

I had an incredible time at Oxford. I would like to thank Dr Eleni Kechagia-Ovseiko, the whole Nuffield College staff and MCR for making the last few years so special. In particular, I thank Marie-Lou Sohnius for being the best friend I could have ever wished for. I also thank: Alex, Antonio, Carla, Cesare, Claire, Dan (x2), Edgar, Ele, Florian, Francesco, Gio, Laura, Marco, Martin, Maksim, Martina, Max, Nef, Naji, Roland, Tiph, Toby, Stefania, and Yang. I am also grateful to some of my best friends at the Oxford Internet Institute: Cailean, Calvin, Felix, Clem,

Hannah, Marta, Manuel, Nayana, and many others. Finally, I wish to thank Chefs Alfonso and Chris, and Ryan for making me feel at home for the last few years.

During my visiting periods and various conferences, I have met great friends. At SciencesPo, among many, I thank: Malo, Luis, Ronja, Selma, and Theo for their friendship. At Yale, I will always treasure my time with Alvise, Annarita, Ellie, Jess, Carlotta, Claudio, Kim, Lore, Nastia, Rapha, Renée, and Sebastian. A special mention goes to Friederike Quint, who has provided invaluable feedback during the writing of this thesis. Other thanks go to Emanuele, Federico, Chris, Gaetano, Paolo, and Tom (x2).

I wish to thank my colleagues at the University of Zurich for welcoming me in a new university and city. I am grateful to Professors Fabrizio Gilardi and Karsten Donnay, as well as Colin, Dror, Emma, Hugo, Jonathan, Maël, Meysam, and Sacha.

The following people have been rather detrimental for the writing of this thesis, but I cannot imagine my life without them. In Oxford, I thank: Ale, Bi, Carlo, Ceci, Chiara, Dante, Davide, Dome, Ele, Filo (x2), Gaia, Giulio, Katia, Mattia, Pietro, Ricky, Ros, Sara, Stefano, and Tamy. Back in Italy, I am grateful to my childhood friends for reminding me that while you may leave Naples, Naples never truly leaves you: Ale, Amy, Anto, Angy, Biagio, Chiara, Co, Corrado, Ely, Dome, Fabry (x2), Faro, Fla, Ghety, Mario, Mita, Peppe, Raff, Ripa, Roby, Sass, and Tux. And around the world, I say thanks to Aleks, Alex, Davide, Carlo, Chantal, Chiara, Chris, Clo, Fla, Fra (x2), Gio, Giulia, Joe, Leo, Lila, Linda, Luke, Mati, Matt, Max, Pawel, Sabina, and Sara.

Athletics has accompanied me for most of my life. I owe so much to this sport. I wish to thank my coaches, Diego Perez, Alan Rowling, and Neil Tunstall, as well as my teammates over the years: Crescenzo, Gabriele, Jamal, Marcello, Meg, Naeem, Francesco, and Raffaele.

Finally, I would like to thank my family: mum, dad, Ciro, and Charles, as well as my aunts, uncles, and cousins for teaching me that deep roots are never reached by the frost. *Questa tesi é dedicata a voi!*

# Contents

**1 Introduction** **2**

1.1 The Rise and Impact of Affective Polarisation in US Politics . . . . . 2

1.2 Literature review . . . . . 4

1.3 Outline . . . . . 14

**2 Measuring Online Polarisation:**

**A Supervised Machine Learning Approach** **17**

2.1 Introduction . . . . . 17

2.2 Measuring Polarisation . . . . . 20

2.3 Data and Methods . . . . . 30

2.4 Results . . . . . 34

2.5 Discussion . . . . . 40

**3 The Temporal Dynamics of Online Polarisation in the US** **44**

3.1 Introduction . . . . . 44

3.2 Who Leads, Who Follows in Spreading Polarisation? . . . . . 46

3.3 Data . . . . . 52

3.4 Methods . . . . . 54

3.5 Results . . . . . 57

3.6 Discussion . . . . . 75

**4 The Network Dynamics of Online Polarisation in the US** **78**

4.1 Introduction . . . . . 78

4.2 The Network Dynamics of Online Polarisation . . . . . 81

4.3 Data . . . . . 86

4.4 Methods . . . . . 89

4.5 Results . . . . . 93

4.6 Robustness checks . . . . . 98

4.7 Discussion . . . . . 104

**5 Conclusion** **109**

**A Appendix** **134**

A.1 Survey Design and Descriptive Statistics . . . . . 134
A.2 Dis/Agreements Across Coders . . . . . 140
A.3 Keyness Analysis by Coders' types . . . . . 142
A.4 Comparison against Simchon et al. (2022) . . . . . 143
A.5 Topic Modelling (Paper 1) . . . . . 148
A.6 Summary Tables of Classifiers' Performances . . . . . 150
A.7 Ego vs Socio-tropic considerations . . . . . 154
A.8 Gender effect on Perception . . . . . 156
A.9 Key Dates for Time-Series Annotations . . . . . 158
A.10 Topic Modelling (Paper 2) . . . . . 160
A.11 Further Details on Time-Series Annotations . . . . . 166
A.12 All VAR Models . . . . . 169
A.13 Full HERGMs . . . . . 175

# List of Figures

1 Comparison of Divisiveness Scores by Posters and Coders . . . . . 37
1 Divisiveness Scores by Tweets' Posters and Coders . . . . . 58
2 Divisiveness Democrat-coded Tweets Over Time . . . . . 61
3 Divisiveness Republican-coded Tweets Over Time . . . . . 63
4 Divisiveness Combined-coded Tweets Over Time . . . . . 65
5 Divisiveness GPT-coded Tweets Over Time . . . . . 67
1 Network visualisation . . . . . 88
2 Network Centrality Measure by Politicians' Party Affiliation . . . . . 94
3 Goodness-of-fit - Geodesic Distance . . . . . 99
4 Goodness-of-fit - In-degree distribution . . . . . 101
5 Goodness-of-fit - Out-degree distribution . . . . . 103
1 Coding Instructions - Example 1 . . . . . 136

2 Coding Instructions - Example 2 . . . 137
3 Coding Instructions - Example 3 . . . 137
4 Example of a coding task . . . 138
5 Examples of tweets with disagreements across coders . . . 140
6 Examples of tweets with agreements across coders . . . 141
7 Keyness Analysis by Coders for Non- vs Divisive rhetoric . . . 142
8 Distribution of Simchon et al. vs Combined Coders' Polarisation Predictions . . . 144
9 Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs Combined Coders 144
10 Distribution of Simchon et al. vs Democrat Coders' Polarisation Predictions . . . 145
11 Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs Democrat Coders 145
12 Distribution of Simchon et al. vs Rep Coders' Polarisation Predictions . . . 146
13 Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs Republican Coders . . . 146
14 Distribution of Simchon et al. vs GPT-4o's Polarisation Predictions . . . 147
15 Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs GPT-4o . . . 147
16 Democrat coders - Summary table for "Divisiveness" . . . 150
17 Republican coders - Summary table for "Divisiveness" . . . 151
18 Combined coders - Summary table for "Divisiveness" . . . 152
19 GPT-4o Coded - Summary table for "Divisiveness" . . . 153
20 Comparison of Divisiveness Scores across Ego (Not-interested) vs Socio-tropic (Interested) Considerations . . . 155
21 Comparison of Divisiveness Scores across Ego (Not-interested) vs Socio-tropic (Interested) Considerations by Partisanship . . . 156
22 Comparison of Divisiveness Scores by Gender . . . 157
23 Comparison of Divisiveness Scores by Gender and Partisanship . . . 157
24 Topic Modelling - Democrat Politicians . . . 161
25 Topic Modelling - Republican Politicians . . . 162
26 Topic Modelling - Democrat-leaning Media . . . 163
27 Topic Modelling - Republican-leaning Media . . . 164
28 Topic Modelling - Public . . . 165

29 Democrat Coded - Cumulative Non-Orth IRF . . . 169
30 Democrat Coded - Cumulative Orth IRF . . . 169
31 Democrat Coded - Non-Cumulative Orth IRF . . . 170
32 Republican Coded - Cumulative Non-Orth IRF . . . 170
33 Republican Coded - Cumulative Orth IRF . . . 171
34 Republican Coded - Non-Cumulative Orth IRF . . . 171
35 Combined Coded - Cumulative Non-Orth IRF . . . 172
36 Combined Coded - Cumulative Orth IRF . . . 172
37 Combined Coded - Non-Cumulative Orth IRF . . . 173
38 GPT Coded - Cumulative Non-Orth IRF . . . 173
39 GPT Coded - Cumulative Orth IRF . . . 174
40 GPT Coded - Non-Cumulative Orth IRF . . . 174

# List of Tables

1 Summary of the coded tweets by tweets' posters and coders' parties . . . 34
2 Summary of intra-coder agreement by party ideology and tweets' posters . . . 35
3 Summary of percentage agreement between inter-coders . . . 36
4 Summary Training Performances by Coders and Specification . . . 39
1 Summary of the Friedrichs et al. (2022)'s Twitter dataset subset under analysis . . . 53
2 Summary of the IRFs output and Granger Causality Tests . . . 69
1 Summary of Key Network Statistics . . . 87
2 HERGMs' within-cluster estimation . . . 96
3 HERGMs' between-cluster estimation . . . 97
4 Multicollinearity checks . . . 99
5 Summary Statistics of Coders . . . 139
6 Summary of the Topics discussed . . . 149
7 Summary of the BERTopic outputs by dataset . . . 160

# 1 Introduction

## 1.1 The Rise and Impact of Affective Polarisation in US Politics

The extensive use of divisive communication and its recent escalation into violence, notably in the US with the January 6th 2021 Capitol Hill Attack, have given urgency to the study of political polarisation and its dynamics (Barberá, 2020; Barberá et al., 2015; 2019; Bail, 2021; Theocharis et al., 2019; Kosmidis and Theocharis, 2020; Bright, 2018). In recent years, polarisation has surged dramatically, having detrimental effects on the quality of US politics. The Pew Research Centre (Kiley, 2017) has conducted a national-representative poll on polarisation over two decades, from 1994 to 2017, measuring the differences in the median ideological estimation (via several 5-point scales) between both Democrats and Republicans. Republicans have been found to be more conservative than 94% of Democrats, while Democrats have been more liberal than 92% of Republicans. Negative opinions towards opposing parties have surged since then, with Republican unfavourable views of Democrats rising from 17% to 43%, and Democrat unfavourable views of Republicans rising from 16% to 38%. More recently, Pew (2023), analysing a survey of 5,115 US adults from June 5 to 11, 2023 and 8,480 US adults from July 10 to July 16, 2023, has found further polarisation, with 72% of Republicans and 63% of Democrats now viewing the opposing party as more immoral than other Americans, against 2016's 47% and 35% respectively. Moreover, open-ended responses show "divisive" as the most frequent feeling about US politics. Focusing on the effects of such polarisation, prominent scholarship has shown that it is detrimental to democratic institutions, propagating intolerance and cynicism, as well as reducing the opportunities for political collaboration and compromise (Hobolt et al. 2020; Layman et al., 2006; Valentino et al., 2008).

While some research suggests that exposure to political negative language, or attacks, can increase voter engagement and interest in politics (e.g., Martin, 2004; Finkel and Geer, 1998), evidence points to negative consequences of aggressive political discourse (Nai and Young, 2024). Studies link negative campaigning to issues like voter demobilisation (Ansolabehere and Iyengar, 1995), decreased trust in politicians (Reiter and Matthes, 2022), lower public mood (Lau et al., 2007; Stevens, 2008), and increased intolerance and cynicism (Yoon et al., 2005; Layman et al. 2006), increased toxicity and incivility in users (e.g., Hopp and Vargo, 2017; Auter and Fine, 2016; Gross

and Johnson, 2016). At the politician level, greater polarisation not only reduced opportunities for collaboration and compromise, but also reinforced incentives to use combative rhetoric to mobilise voters despite the risks of further eroding democratic norms (MacKuen et al. 2010; Nai and Young, 2024).

Previous work has also showed that aggressive behaviour by political elites tends to resonate specifically with certain segments of the population, such as voters who exhibit aggressive tendencies themselves (Kalmoe, 2014; Nai and Young, 2024), have higher levels of negative personality traits (Nai and Maier, 2021), or possess a strong orientation toward conflict (Sydnor, 2019). Nai and Maier (2021) study how individual differences influence responses to attack politics through two online experiments with US participants on Amazon Mechanical Turk (N=1,408 and N=1,081). The first assesses how personality traits affect reactions to character versus policy attacks, while the second study compares civil and uncivil attacks, exploring moderation strategies, negativity tolerance, and conflict avoidance. Nai and Maier (2021) find that responses to negativity and incivility vary significantly. Individuals high in conflict avoidance view the sponsor more negatively after exposure to attacks, while those high in psychopathy have a more negative opinion of the target following character attacks. Thus, aggressive political campaigns resonate with some voters but are strongly rejected by others. In more recent work, Nai and Maier (2023) show that differences in individual perceptions contribute significantly to the link between negativity and affective polarisation. Therefore, to develop a more comprehensive model, the authors argue that it is essential to examine the sources of these perceived differences, which may stem from individual traits, as well as broader social and cultural norms.

Research on the interaction between media, public opinion, and polarisation offers nuanced insights into how exposure to political information influences public attitudes and preferences (Garrett et al., 2014; Wlezien, 2024; Wlezien and Soroka, 2024). Garrett et al. (2014) analyse post-election surveys from the US and Israel to study affective polarisation, defined by net favourability toward in-party vs. out-party, as influenced by exposure to ideologically biased news. Their findings reveal that increased exposure to pro-attitudinal news, which often portrays the in-group positively and the out-group negatively, correlates with higher levels of affective polarisation in both countries. Conversely, greater exposure to counter-attitudinal news, which likely includes criticisms of the in-group, is linked to a reduction in affective polarisation. The authors attribute this effect to

cross-pressure and "contact" theories (e.g., Mutz, 2002; Mutz and Mondak, 2006), suggesting that exposure to opposing views can lead to more moderate perspectives. However, these findings are based on the assumption that congruent and incongruent political news contain partisan attacks toward and from the out-group, respectively. On the role of the media, Wlezien (2024) and Wlezien and Soroka (2024) produce strong evidence to argue that media content typically reflects, rather than shapes, public and policy preferences, amplifying existing trends in areas like defence, welfare, and health care. Such investigations will be further examined throughout the thesis.

Given these complexities, a comprehensive understanding of polarisation requires investigating not only how elites and media deploy divisive rhetoric, but also how individuals perceive and respond to such communication. This thesis addresses these gaps by examining the mechanisms through which polarised language is produced, disseminated, and interpreted in the digital public sphere, with a particular focus on the subjective and context-dependent nature of polarisation.

## 1.2 Literature review

This literature review is organised into five sections to provide a comprehensive foundation for the thesis. First, I present a definition of online polarisation, which will serve as a conceptual framework throughout the following chapters. Second, I examine previous research on the measurement of polarisation, highlighting the various methodologies and metrics employed in the field. Third, I review the literature concerning the temporal dynamics of polarisation, focusing on how polarisation evolves across users and over time. Fourth, I analyse scholarly work on the network dynamics of polarisation, with particular attention to how social connections and structures influence polarised discourse. Finally, I explain the rationale behind my case selection, specifically my focus on online polarisation within the US as manifested on Twitter[1].

### 1.2.1 Defining Polarisation

Before engaging in my analytical work, I first discuss the different definitions of affective polarisation. The theory of affective polarisation states that, in political discussions, individuals have positive attitudes towards their in-group members and negative ones towards the out-group (see,

[1]Throughout this thesis, I refer to the platform "X" by its former name, Twitter, to comply with previous and cited literature and with the period of data collection.

Iyengar et al., 2012; Iyengar and Westwood, 2015; Mason, 2015; 2018). In other words, positive emotional attachments are used towards co-partisans and greater hostility is used towards opposing partisans. Prominent research places polarisation within political psychology alongside social identity theory, where partisanship becomes the most salient social and political identity (Luttig, 2018; Mason, 2015; 2016; Iyengar, et al., 2012). Affective polarisation is a branch of the comparative politics literature that focuses on explaining political division (Garrett and Bankert, 2018; Mason, 2013; 2015; Lelkes and Westwood, 2017). It studies the rising levels of animosity towards opposing groups and favouritism towards allied or coalition groups (Reiljan, 2016; Iyengar et al., 2012; 2019; Lelkes and Westwood, 2017; Mason, 2013; 2015). Affective polarisation argues that partisanship can be seen as a group identity which, in turn, generates in-group bias as well as discrimination against opposing out-groups (Iyengar et al., 2012; Miller and Conover, 2015; Iyengar and Westwood, 2015). While this theory seeks to study the emotional distance between partisans, it is the question of how, and under what conditions, this occurs and what motivates the identification of the dynamics of emotional group-based polarisation. On the emotional side, "negative animosity between parties" fuels polarisation, where animosity refers to negatively-charged language against opponents (Iyengar et al., 2012; 2015; 2019: 129).

Building on these premises and to study divisive communication, this thesis employs a minimal definition of polarisation, that is as positive emotional attachments towards supporters of one's party (co-partisans) and greater hostility towards members of the opposing group/party (opposing partisans) (Iyengar et al., 2012; Iyengar and Westwood, 2015; Mason, 2015; 2018). Consistent with Fletcher et al. (2019), I do not expect online polarisation to differ significantly from offline polarisation. Their study across 12 countries, including the US, Europe, and Australia, shows minimal difference between the two. However, the advantage of studying online data remains significant as it offers insights that are more granular, extensive and often more precise than surveys given the large datasets that track a range of topics over time.

It is important to discuss what this conceptualisation of polarisation leaves out. Alternative studies often investigate polarisation in terms of issue positioning, or issue polarisation, considering the relationship between individual-level socio-economic preferences and/or identity, and polarisation (Hobolt and De Vries, 2016; De Groot, 1974). In this debate, Barberá (2020) distinguishes between the divergence in political views, or issue polarisation, and dislike for the partisan out-group, known

as affective polarisation. Research has found little or no correlation between issue polarisation and affective polarisation (Iyengar et al., 2019; Levendusky and Malhotra, 2016). Earlier work suggests that individuals do not experience cognitive biases when processing political messages they receive; instead, they simply update their stances based on the received information (De Groot, 1974). Later work on partisan-motivated reasoning finds the opposite (Lodge and Taber, 2013; Redlawsk, 2002). When a citizen is presented with negative information about a candidate they evaluate positively, their support for that candidate may increase (instead of decrease) due to affective biases (Redlawsk, 2002). Such biased information processing appears to have a strong impact on polarisation in the US. For example, even when Democrat and Republican voters may find agreement on policy positions, they tend to dislike and distrust each other more, and often perceive their ideological distance to increase (Mason, 2018). Measures such as DW-NOMINATE, which estimate legislators' ideological positions based on their roll-call voting records, have been widely used to track the increasing ideological distance between parties in the US Congress over time, illustrating the empirical rise of polarisation since the 1970s (Hare and Poole, 2014).

### 1.2.2 Measuring Polarisation

A range of methodologies has been employed to assess political polarisation, with survey experiments, network-based analyses, and dictionary-driven NLP techniques among the most prominent. Although survey experiments are widely used, their findings are often constrained by the artificiality of experimental environments, which can limit the generalisability of results to actual social media behaviour (Asimovic et al., 2021; Kubin and von Sikorski, 2021; Reisman and Dvir-Gvirsman, 2024; Liang and Zhang, 2021). For example, Liang and Zhang (2021) have conducted two population-based online survey experiments in Hong Kong to examine how individuals perceive incivility in political disagreement comments. Their findings showed that people perceive a higher degree of incivility in disagreement comments directed at their own in-group compared to out-group members, and that exposure to such comments influences affective polarisation and willingness to participate in politics indirectly through perceived incivility. In response, field experiments have emerged, leveraging interventions in participants' genuine social media environments to enhance ecological validity (Guess et al., 2023a; 2023b; Levy, 2021). A notable example is Guess et al. (2023a), who conducted a large-scale field experiment during the 2020 US election

by assigning Facebook and Instagram users to either algorithmic or chronological feeds. While the chronological feed changed what users saw and reduced their time on the platforms, it did not significantly affect levels of polarisation, political knowledge, or key political attitudes. However, despite these advances, computational methods remain indispensable for analysing polarisation at scale in real-world contexts.

Social network analysis, as demonstrated by Barberá et al. (2015), has illuminated the ways in which ideological clustering shapes information flows on platforms like Twitter. By inferring users' ideological leanings from their network connections, such research has shown that political discussion is often dominated by exchanges within ideologically similar groups. Nevertheless, this approach is limited by its narrow topical and temporal scope, which may fail to capture the shifting and multifaceted nature of online discourse. Further, the exclusion of users who do not fit certain criteria introduces selection bias, and a focus on network structure alone overlooks the substance and sentiment of the communications being analysed. Dictionary-based NLP approaches, such as those developed by Simchon et al. (2022), have sought to systematically identify polarised language through curated word lists. While these tools can effectively distinguish partisan communication and are adaptable across platforms, their coverage is inherently limited by the static nature of dictionaries and the rapid evolution of online vernacular. Additionally, using metrics like word frequency or message length as indicators of polarisation can oversimplify the complex ways in which divisive rhetoric is constructed and interpreted in digital spaces.

A fundamental shortcoming of both network and dictionary-based frameworks is their implicit assumption that polarisation is perceived similarly by all observers. In practice, the process of annotating or coding content is susceptible to the subjective biases of individuals, particularly those related to partisan identity (Ashokkumar et al., 2020; Van Bavel and Pereira, 2018; Kunda, 1990; Lord et al., 1979; Taber and Lodge, 2006; Appel et al., 2023; Ecker et al., 2022; Pennycook and Rand, 2019; Pennycook et al., 2021). Motivated reasoning, confirmation bias, and lapses in attention can all influence how coders interpret and label divisive content, potentially distorting measurement outcomes.

To address scalability and reliability, crowdsourcing has been adopted as an alternative to traditional expert annotation, with the "wisdom of the crowds" approach offering notable flexibility and

efficiency (Benoit et al., 2016). However, the effects of coder diversity—especially in terms of ideological orientation—remain underexplored. Recent studies indicate that partisanship shapes how individuals perceive and rate divisiveness, with coders tending to minimise polarisation within their own group and amplify it in the opposing camp (Appel et al., 2023; Kozyreva et al., 2023; Williams et al., 2025). This has significant implications for the validity of annotated corpora and the models trained on them. Notably, Williams et al. (2025) demonstrate that annotator demographics—such as gender and political identity—systematically influence how political texts and images are labelled. The authors analyse two cases: one involving image annotation in the US and another with text annotation in the Netherlands. In both, they find that annotator identities lead to significant differences in labelling, even for seemingly objective tasks. The study shows that these biases can affect research validity and the outcomes of models trained on such data.

The advent of generative AI, and in particular large language models (LLMs), has introduced new possibilities for text labelling. These models have demonstrated performance comparable to, and in some cases exceeding, that of human annotators on specific tasks (Gilardi et al., 2023; Yang and Menczer, 2023; Ziems et al., 2023; Törnberg, 2024). In this regard, LLMs can offer consistent and scalable labelling, but their tendency to favour neutral responses means that human oversight remains important for ensuring nuanced and contextually appropriate classification (Huang et al., 2024).

In summary, this first section reveals several persistent challenges in measuring polarisation in textual data. There is a pressing need for more sophisticated tools that can capture the full range of divisive language, leveraging advanced machine learning techniques. The influence of coder ideology on annotation outcomes warrants systematic scrutiny, including direct comparisons with AI-generated labels. Finally, greater attention to coder heterogeneity promises to enhance the robustness and validity of polarisation measurement, enabling a more detailed and dynamic understanding of public discourse. This thesis addresses these issues by developing and evaluating transformer-based classifiers trained on both human and AI-coded data, with a particular emphasis on the interplay between coder's partisanship and measurement outcomes.

### 1.2.3 Temporal Dynamics of Polarisation

The surge of polarisation in digital political communication has prompted scholars to reconsider the roles and interactions of various actors on social media platforms. While much attention has been paid to the mechanisms by which echo chambers and algorithmic curation reinforce ideological divides (Bail, 2021; Barberá, 2020; Hobolt et al., 2024; Ross Arguedas et al., 2022; Sunstein, 2018). Jamieson and Capella (2008: 76) define the former as "a bounded, enclosed media space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal". However, the precise origins and escalation of polarising language among politicians, journalists, and ordinary users remain less clearly mapped. Social media's architecture fosters environments where individuals gravitate toward like-minded communities, amplifying their perspectives and potentially muting dissent (Pariser, 2011; Sunstein, 2018). Political elites, in turn, may strategically deploy antagonistic rhetoric to energise supporters, while grassroots users—often embedded in homogenous networks—can both mirror and initiate divisive exchanges (Barberá et al., 2019; Theocharis et al., 2020). Recent scholarship underscores the need to disentangle the temporal and actor-driven dynamics that underpin online polarisation, especially during salient events such as elections or crises (Harteveld, 2021; Iyengar and Wagner, 2024; Röllicke, 2023; Stavrakakis, 2018; Westphal, 2024).

A foundational distinction in this literature is between polarisation that emerges horizontally—among citizens themselves—and vertical polarisation, which unfolds between the public and elites, either as bottom-up pressure or top-down mobilisation (Harteveld, 2021; Harteveld et al., 2022; Röllicke, 2023; Westphal, 2024). The media, meanwhile, often serve as an intermediary, amplifying signals that originate elsewhere (Wlezien, 2024; Wlezien and Soroka, 2024). Despite these insights, questions persist about which actors most frequently initiate polarising rhetoric and the conditions that prompt its intensification.

Empirical research indicates that politicians are often highly responsive to the preferences and emotions of their core supporters, particularly during high-stakes periods such as electoral campaigns (Barberá et al., 2019; Gerber and Morton, 1998; Holbrook and McClurg, 2005; Egan, 2013; Kastellec et al., 2015; Neundorf and Adams, 2018). Rather than acting as agenda-setters, political elites frequently adapt their messaging to reflect the sentiments of their base, using emotional ap-

peals to reinforce group identity and mobilise participation (Crabtree et al., 2020; Kosmidis et al., 2019; Müller, 2022; Valentim and Widmann, 2021; Widmann, 2022). The often-rapid feedback mechanisms of social media intensify this dynamic, allowing politicians to adjust their rhetoric in near real-time according to public sentiment (Hawkins et al., 2020). However, research suggests that politicians primarily use these platforms for broadcasting rather than direct engagement, in part to avoid negative interactions or loss of narrative control (Stromer-Galley, 2000; Ward and Lusoli, 2005; Tromble, 2018). Nonetheless, evidence shows that public expressions online can influence elite responses, indicating a more reciprocal relationship than previously assumed (Tromble, 2018).

The media's function in this ecosystem is often characterised as reactive or catalytic rather than initiatory. Rather than directly shaping public attitudes, news outlets tend to reflect and amplify the prevailing sentiments and actions of both the public and politicians (Wlezien, 2024; Wlezien and Soroka, 2024). This feedback loop suggests that media coverage tends to mirror, and sometimes intensify, existing divisions rather than create new ones.

Cross-partisan exchanges—interactions between users with differing political allegiances—are another focal point of recent research. While such interactions can, under certain conditions, foster moderation and reduce polarisation (Holbert et al., 2010; Bright, 2018; Hong and Kim, 2016), they are less frequent among users at the ideological extremes, who tend to remain insulated within homogeneous communities (Bright et al., 2018; Conover et al., 2011, 2012; Levendusky, 2013). Social media's design can both facilitate and constrain these encounters, and the pursuit of engagement—through likes, shares, and retweets—may further incentivise the production and dissemination of polarising content (Bail, 2021; Lin et al., 2023). Experimental evidence further demonstrates that the composition of discussion groups—whether ideologically homogeneous or mixed—can significantly shape both affective and policy polarisation (Hobolt et al., 2024). Echo chambers tend to reinforce polarisation, while mixed-partisan groups may, depending on context, foster more moderate discourse.

Despite these advances, significant gaps remain in our understanding of how polarisation emerges and evolves over time, the relative influence of different actors, and the specific conditions under which cross-partisan interactions either heighten or mitigate divisions. This thesis seeks to address these gaps by leveraging supervised machine learning classifiers—trained on both human and AI-

coded data—to systematically measure polarisation in a comprehensive dataset of tweets from political elites, media organisations, and the general public. By analysing the timing, directionality, and actor-specific contributions to polarising language, this research aims to clarify the leadership and followership dynamics in the online polarisation process, as well as the role of partisanship and media amplification in shaping these trends.

### 1.2.4 Network Dynamics of Polarisation

The transformation of political communication by platforms like Twitter has fundamentally altered how elites, media, and the public interact, making the study of online polarisation indispensable for understanding contemporary politics. Twitter, in particular, has become a central arena for both the dissemination of political information and the performance of partisan identities, providing politicians with direct channels to their supporters and bypassing traditional media gatekeepers (Benkler, 2007; Yoon and Park, 2014; Smith, 2011). This digital shift has not only facilitated the rapid spread of information but has also heightened the visibility and immediacy of political alliances and conflicts (Tucker et al., 2017; Jungherr, 2016; Enli and Skogerbø, 2013; Theocharis et al., 2016). As a result, the boundaries between formal and informal political discourse have become increasingly blurred, with social media serving as both a public stage and a battleground for competing narratives (Tucker et al., 2018).

Within this environment, the interactive features of Twitter—such as retweets, mentions, and replies—play a crucial role in shaping the contours of political networks. Retweeting and following are commonly used to signal support and reinforce intra-party cohesion, while mentions can function as tools for both alliance-building and public confrontation (Avalle et al., 2024; Arnaboldi et al., 2022; Cherepnalkoski et al., 2016; Del Vicario et al., 2016; Ogawa et al., 2013). Despite the platform's potential for exposing users to a diversity of viewpoints (Barberá et al., 2015; Colleoni et al., 2014; Vaccari et al., 2016), empirical research consistently finds that political elites gravitate toward ideologically similar clusters, with cross-party engagement often taking the form of antagonistic exchanges rather than constructive dialogue (Vergeer, 2020; Boutet et al., 2012; Himelboim et al., 2013).

Network analysis has proven invaluable for mapping these patterns, revealing how structural features like reciprocity, clustering, and political homophily shape the formation and persistence of partisan

communities (Esteve Del Valle and Borge Bravo, 2018; Esteve Del Valle et al., 2022; Song, 2015; Fraxanet et al., 2024; Intal et al., 2021; Capozzi et al., 2023; Quelle and Bovet, 2025). While retweets and follows tend to reinforce ideological boundaries, mentions can act as bridges between camps—though these bridges are frequently used for rhetorical conflict rather than consensus-seeking (Conover et al., 2011; Esteve Del Valle and Borge Bravo, 2018a; Esteve Del Valle et al., 2020; Yoon and Park, 2014).

The content of these interactions is equally important for understanding polarisation. Research shows that cross-party mentions and interactions are often marked by negative or divisive language, reflecting affective polarisation's core dynamic: strong positive sentiment toward in-group members and pronounced hostility toward out-groups (Iyengar et al., 2012; Iyengar and Westwood, 2015; Mason, 2015, 2018; Avalle et al., 2024; Del Vicario et al., 2016; Qiu et al., 2019; Hegelich and Shahrezaye, 2015). Experimental evidence further demonstrates that exposure to counter-attitudinal content on social media can intensify, rather than mitigate, partisan divisions, particularly among those with strong ideological commitments (Bail et al., 2018; Nyhan and Reifler, 2010; Taber and Lodge, 2006; Bail, 2015). Despite the promise of social media to foster pluralism and cross-cutting dialogue, the prevailing evidence suggests that elite networks are largely shaped by partisan clustering and conflict-driven engagement. While mentions and replies offer opportunities for cross-party contact, these interactions are more often vehicles for contestation than for deliberation or compromise (Avalle et al., 2024; Del Vicario et al., 2016; Qiu et al., 2019; Hegelich and Shahrezaye, 2015). This reinforces the need for research that integrates both network-structural and content-based approaches to fully capture the dynamics of polarisation among political elites.

This thesis responds to these gaps by combining advanced social network analysis with supervised machine learning techniques to examine both the structural and discursive dimensions of elite polarisation on Twitter. By systematically analysing the interplay between network position, interaction patterns, and the tone of political communication, this research aims to provide a more nuanced understanding of how digital platforms both reflect and reinforce partisan divisions in contemporary political life.

### 1.2.5 Case Study

This thesis focuses on Twitter and, specifically, US politics to study polarisation. Research on Twitter communication is extensive (Barberá et al., 2014; 2016; 2019; Kosmidis and Theocharis, 2020; Bright, 2018; Bright et al., 2018), as the platform offers rich textual data from a diverse user base, including celebrities, politicians, and everyday citizens. Its character limit encourages concise expression, making the data easy to analyse. Additionally, Twitter provides valuable metadata, such as engagement metrics and follower information, making it an ideal platform for studying online polarisation. One way to study polarisation is via negative and positive discussions on this platform. Examples of negative language are represented by tweets that display high levels of negative emotions, such as anger and disgust (Kosmidis and Theocharis, 2020). In contrast, positive language is associated with positively-charged emotions, such as joy and trust (ibidem). Some researchers have considered negatively-charged language to be "milder" when mocking or insulting (Anderson et al., 2013; Borah, 2012; Massaro and Stryker, 2012; Kosmidis and Theocharis, 2020), while other scholars have studied extremely negative language in the instances of homophobia, racism and sexism (Theocharis et al., 2016; Kosmidis and Theocharis, 2020). Polarised language is thus commonly perceived as a continuum in which positive language is placed at one end while mild cases are in the centre and stronger cases are at the extreme (Sydnor, 2019; Kosmidis and Theocharis, 2020). Barberá's (2014) topic modelling research studies political communication by reference to political events (e.g., elections, protests), policy proposals (e.g., same-sex marriage), and economic issues (e.g., economic reforms, redistribution preferences). Non-political topics can be used as a placebo to test how polarised language may affect other domains of life (ibidem).

More recent scholarship argues that political actors are moving online, especially following Obama's 2008 and 2012 successful campaigns, and are becoming more engaged in online discussions (Barberá et al., 2019; McLoughlin and Ward, 2020; Theocharis et al., 2020; Tromble, 2018). On the supply-side, research shows that political elites widely utilise social media websites (e.g., Twitter) to share/promote their political brand with the intent of mobilising and creating political support (Kosmidis and Theocharis, 2020; Theocharis et al., 2020). On the demand-side, public users have increasingly used Twitter for news consumption and interaction with politicians, who used to be hard-to-reach individuals (Kosmidis and Theocharis, 2020; Newman et al. 2017).

The US is an ideal case for examining polarisation through social media, particularly Twitter, due to several unique factors. First, in the US there is an extensive and influential social media presence on Twitter, where the platform plays a central role in public discourse and political communication (e.g., Barberá et al., 2015; Bode and Vraga, 2018; Bail et al., 2021). This platform's reach and popularity provide rich data for analysing how language and sentiment contribute to societal divides. Specifically, according to Pew Research Center (2024), the percentage of US adults who reported ever using Twitter rose gradually from 22% in 2015 to 31% in 2020, with an average of 27%. Looking at the period under analysis in this thesis, Pew found that 22% of US adults used Twitter in 2015, 25% in 2016, 29% in 2018, 28% in 2019, and 31% in 2020, indicating a steady increase in Twitter adoption among American adults. Second, the US political system is majoritarian, typically characterised by two dominant parties, making it conducive to a binary classification approach for studying polarisation (e.g., Abramowitz and Saunders, 2008). This structure allows for clearer distinctions between opposing viewpoints, which simplifies the analysis of polarised language. Additionally, the English language, predominantly used in US social media, is the basis for many natural language processing tools, including the BERTweet model, which is designed specifically for Twitter data (Nguyen, Vu and Tuan, 2020). Paper 1 includes a detailed explanation of its training and employment in this thesis. Lastly, the US serves as a global reference point for research on digital political communication, with a substantial body of comparative and longitudinal studies documenting the evolution of polarisation on Twitter across multiple election cycles (e.g., Barberá, 2014; Barberá et al., 2015; Barberá, 2020; Bail et al., 2021; Tucker et al., 2018). This allows the thesis to build on robust empirical foundations while contributing new insights into the mechanisms and consequences of polarised language in digital environments. In the conclusion to the thesis, the recent changes to Twitter, as well as their impact on the findings, will be discussed.

## 1.3 Outline

This thesis addresses my research questions through a structure of three complementary papers. It examines the mechanisms of affective polarisation within the US public discourse online, offering innovations in measuring polarisation in textual data, focusing on interactions between politicians, news media, and the general public. After completing an overall introduction and literature review, this thesis proceeds as follows.

The first paper tackles the methodological challenge of detecting polarising language in social media texts. By developing and validating supervised machine learning classifiers based on transformer models, I systematically compare the influence of coder ideology and generative AI on the annotation and detection of divisive content. The analysis reveals that human coders' political affiliations significantly shape their evaluations, with Democrats rating tweets as more divisive than Republicans or AI models, especially when assessing out-group messages. Notably, models trained on a combination of coder perspectives outperform those based on single partisan or AI-only annotations, underscoring the value of coder diversity for balanced measurement. These classifiers not only improve accuracy in identifying polarisation but are also made publicly available for broader research applications.

The second paper investigates the temporal and directional dynamics of affective polarisation among US politicians, news media, and the public. Leveraging the classifiers developed in the first study, I analyse a large-scale dataset of tweets from 2015 to 2020 to disentangle the reciprocal influences shaping polarising rhetoric. The findings challenge linear models of influence: vector-autoregressive analyses demonstrate that politicians, media, and public users are locked in a dynamic feedback loop, with each actor alternately leading and responding to shifts in divisive language. The media's catalytic role and the prevalence of cross-polarisation interactions reveal a complex ecosystem in which polarisation is both produced and sustained by multiple actors, rather than imposed from a single source.

The third paper turns to the structural foundations of elite polarisation on Twitter, examining how patterns of interaction and network architecture reinforce partisan divisions. Using hierarchical exponential random graph models, I map the relational dynamics among US politicians. The paper finds that party-based homophily dominates network formation. Democrat politicians hold stronger connections to other powerful nodes in the network than Republicans, showing different communication strategies. The most divisive language is concentrated in the rare ties bridging partisan clusters, highlighting the joint role of network structure and rhetoric in entrenching polarisation.

In sum, this thesis advances the study of political polarisation by combining theoretical and methodological innovations with multi-level empirical analysis. By developing new tools for detecting divisive language, tracing its evolution across actors and time, and mapping the network structures

that sustain partisan divides, the research demonstrates that polarisation is a dynamic and deeply embedded feature of online political discourse. While the findings highlight the persistent influence of partisan identity and the reinforcing effects of digital networks, they also provide a flexible framework for future research and practical interventions. Despite certain limitations in scope and generalisability, which are discussed in the individual chapters and conclusions, this work offers both theoretical insight and practical guidance for understanding and addressing the challenges of polarisation in the digital public sphere.

# 2 Measuring Online Polarisation: A Supervised Machine Learning Approach

**Abstract**

Social media interactions are often plagued with polarising, or divisive, language. However, very few studies have been able to adequately detect and measure polarisation at scale and over time. This paper introduces a methodological solution to this problem, using supervised machine learning that leverages transformer models. I develop novel classifiers to detect polarising language in social media texts along a continuous scale, considering coders' political affiliations, as well as benchmarking it against generative artifical intelligence (AI). This approach not only generates accurate algorithms, but also explores human coders' evaluations, and potential deviations between human and machine perspectives. Results reveal that Democrat coders rate tweets as more divisive than Republicans, Combined coders, and GPT-4o. Among humans, Democrats assign higher divisiveness scores to Republican posters than Republicans, reinforcing partisan influences on coding. GPT-4o consistently assigns the lowest divisiveness scores across all groups, significantly diverging from human assessments. When training algorithms, the best-performing model is based on Combined coders, followed by the Republican, Democrat and GPT-4o-trained models. These findings highlight the value of integrating diverse perspectives, as partisan biases may counterbalance each other to yield more balanced assessments. While GPT-4o demonstrates consistency, it does not surpass human coders in classification accuracy, underscoring the continued importance of human judgment in detecting divisive content. The classifiers developed in this study will be publicly available and adaptable to various social media platforms and contexts.

**Keywords:** American Politics, Polarisation, Social Media, Machine Learning, Transformers.

## 2.1 Introduction

The exploration of online political polarisation has become increasingly urgent, particularly in light of recent events that have highlighted the divisive nature of online discourse. Social media platforms empower users to create affective narratives around significant events through hashtags,

allowing direct engagement with political elites and their statements in both civil and uncivil ways (Papacharissi, 2015; Theocharis et al., 2016; McLoughlin and Ward, 2017). However, the rise of polarising language in these interactions has escalated, culminating in real-world violence, such as the January 6th attack on the US Capitol (Barberá et al., 2015; Barberá, 2020; Bail et al., 2021). To understand the mechanisms behind this polarisation, it is essential to examine how political elites utilise divisive language to galvanise their supporters while antagonising their opponents, along with the role of news media in shaping this communication. However, the perception of polarisation is inherently subjective, influenced by individual ideological leanings and personal experiences. What one person views as a polarising statement may be perceived differently by another, depending on their political affiliation, cultural background, and personal biases.

This paper adopts a minimal definition of polarisation, emphasising positive affective attachments to co-partisans and heightened hostility towards opposing partisans (Iyengar et al., 2012; Iyengar and Westwood, 2015; Mason, 2015; 2018). By focusing on online polarisation, it draws parallels to offline phenomena while acknowledging the unique characteristics of digital interactions (Fletcher et al., 2019). The examination includes an analysis of negative language in social media posts—often marked by emotions like anger and disgust—against positive language characterised by joy and trust (Kosmidis and Theocharis, 2020).

Crucially, existing computational measurement strategies, which mainly utilise network (e.g., Barberá et al., 2015; Guerra et al., 2021; Marozzo and Bessi, 2018; Peralta et al., 2024) and dictionary-based approaches (e.g., Simchon et al. 2022), present some methodological limitations in assuming that polarisation is perceived uniformly, which I seek to address. This subjectivity extends to the measurement and analysis of polarisation itself. Researchers and coders tasked with labelling content may inadvertently introduce their own biases into the process (Appel et al., 2023; Benoit et al., 2016; Carlson and Montgomery, 2017; Cheng et al., 2020; Huddy et al., 2023; Kozyreva et al., 2023; Lindner and Nosek, 2009). Such differences may be due to partisans' perception of the truth and their interest in preserving their own identity (Ashokkumar et al., 2020; Van Bavel and Pereira, 2018; Kunda, 1990; Lord et al., 1979), which in turn instruct partisans' motivated reasoning (e.g., Taber and Lodge, 2006), prior attitude effect, and confirmation bias, as well as more general cognitive mechanisms, like inattention (Appel et al., 2023; Ecker et al., 2022; Pennycook and Rand, 2019; Pennycook et al., 2021).

Based on these premises, following previous literature (Appel et al., 2023; Kozyreva et al., 2023; and extending Benoit et al. (2016)'s work, I argue that coders' in-group identities may significantly influence their evaluation and labelling of public discourse. Hence, I propose a group-based "wisdom of the crowds" strategy could improve the measurement of polarisation. Consequently, in the US context, I expect that Republican and Democrat coders will have differing perceptions of what constitutes polarisation. For instance, Republican coders may overlook divisiveness within their party's tweets, while Democrats may be reluctant to label divisive posts from their fellows. This raises important questions about whether such biases extend beyond partisanship, especially when comparing human assessments with those made by generative AI. Accounting for heterogeneity of the coding crowd may enhance the accuracy and robustness of classifiers for identifying and measuring polarisation in public discourse. Moreover, it allows for a direct measurement of partisan specific polarisation; a useful tool for anyone interested in how elites interact with public opinion and what the consequences of these interactions might be. By considering a wider range of partisan perspectives, this approach seeks to mitigate individual coder biases and capture more fine-grained expressions of polarisation. Additionally, diverse input may improve the model's ability to detect more subtle shifts in tone, context, and sentiment, leading to a more comprehensive and dynamic understanding of polarisation as it evolves across groups and over time. Hence, the research question that I seek to investigate is the following:

> *How do coders' partisanship and the use of generative AI impact the measurement of polarisation in social media text analysis?*

My investigation is based on the comprehensive dataset of US geotagged tweets posted between January 2015 and December 2020, developed by Friedrich, Stoehr and Formisano (2023). Using survey respondents and generative AI, I train four supervised machine learning classifiers, via transformer models, to annotate text on containing polarising language along a continuous scale. To fully understand the classifier's decisions, I analyse coders' responses in relation to their ideological positioning (i.e., whether they identify as Republican or Democrat). Finally, as a benchmark, I repeat the tweet-coding process, using a pre-trained generative artificial intelligence model, OpenAI's GPT-4o, and I compare human against machine labelling.

The results reveal that this approach not only generates accurate algorithms, but also explores human

coders' evaluations, and potential deviations between human and machine perspectives. Results show that Democrat coders rate tweets as more divisive than Republicans, Combined coders, and GPT-4o. Among humans, Democrats assign higher divisiveness scores to Republican posters than Republicans, reinforcing partisan influences on coding. GPT-4o consistently assigns the lowest divisiveness scores across all groups, significantly diverging from human assessments. When training algorithms, the best-performing model is based on Combined coders, followed by the Republican, Democrat and GPT-4o-trained models. This paper makes significant contributions to polarisation research. First, I develop and publicly release advanced machine learning classifiers that accurately identify divisive language on a continuous scale, improving existing measurement tools. Second, I assess the classifiers' outputs by examining how coders' partisan identities shape their decisions, using state-of-the-art generative AI as a benchmark. Such a combination of algorithm development and coders' partisanship evaluation is a novel approach in polarisation research, and it can be applied at scale, using different languages and text types, as well as over time.

## 2.2 Measuring Polarisation

Social media sites allow people to build affective narratives about a distinct event around specific hashtags (Papacharissi, 2015), directly respond to the political elite making a particular statement in a civil or uncivil manner (Theocharis et al., 2016; McLoughlin and Ward, 2017), and quickly engage with supporters of the other side (Bail et al., 2018; Barberá et al., 2015). However, social media interactions are often plagued with polarising language, and their recent escalation into violence, such as the January 6th attack on the Capitol in Washington, has created an urgency to study political polarisation and its dynamics (Barberá, 2020; Barberá et al., 2015; 2019; Bail, 2021; Theocharis et al., 2019; Kosmidis and Theocharis, 2020; Bright, 2018). A plausible explanation behind these patterns is one in which elites use divisive language to widen their audience (e.g., Lawrence and Boydstun, 2017). By being divisive, elites manage to provoke their ideological opponents and, consequently, encourage their in-group to defend them, triggering uncivil interactions with out-groups.

As previously discussed, this paper employs a minimal definition of polarisation, that is positive emotional attachments towards supporters of one's party (co-partisans) and greater hostility towards members of the opposing group/party (opposing partisans) (Iyengar et al., 2012; Iyengar and

Westwood, 2015; Mason, 2015; 2018). Although my work concentrates on online polarisation, I believe that Iyengar et al.'s definition is well-suited for my research, as it effectively captures the essential dynamics of emotional attachments and hostility between partisans. These dynamics are also evident in online interactions, reflecting the substantial similarities between online and offline polarisation (Fletcher et al., 2019). In terms of such polarised communication, examples of negative language are represented by social media posts that display high levels of negative emotions such as anger and disgust (Kosmidis and Theocharis, 2020). In contrast, positive language is associated with positively-charged emotions, such as joy and trust (ibidem). Some researchers have considered negatively-charged language to be "milder" when mocking or insulting (Anderson et al., 2013; Borah, 2012; Massaro and Stryker, 2012), while other scholars have studied extremely negative language in the instances of homophobia, racism and sexism (Theocharis et al., 2016; Kosmidis and Theocharis, 2020). Polarised language is thus commonly perceived as a continuum in which positive language is placed at one end while mild cases are in the centre and stronger cases are at the extreme (Sydnor, 2019; Kosmidis and Theocharis, 2020).

This review proceeds as follows. First, I review two seminal works that attempt to measure polarisation via network and dictionary-based approaches. Second, I discuss the importance of coders' partisanship perceptions. Third, I discuss how pre-trained generative AI can aid in such classification tasks. Finally, I summarise my contributions.

### 2.2.1 Measuring Online Polarisation

To frame this discussion, I begin by reviewing common methodologies for measuring polarisation, particularly survey experiments, which are widely used despite limitations in capturing authentic social media dynamics. I then shift to two significant computational approaches in polarisation research: social network analysis, exemplified by Barberá et al. (2015), and natural language processing, as demonstrated in Simchon et al. (2022). These computational methods each stem from distinct research traditions, with network analysis offering robust insights into social connections, while natural language processing techniques focus on classifying textual polarisation, often centring on partisanship rather than divisive rhetoric.

Survey experiments are a popular research method to measure polarisation (e.g., Asimovic et al., 2021; Kubin and von Sikorski, 2021; Reisman and Dvir-Gvirsman, 2024). For instance, Liang

and Zhang (2021) exposed participants of different political affiliations to either civil or uncivil comments in a news story, assessing affective polarisation through questionnaires. However, such experimental designs may struggle with external validity, as their artificial, controlled setups may not reflect natural social media interactions. In an attempt to overcome such limitations, recent technological advances have enabled new experimental designs. For example, Guess et al. (2023a; 2023b) and Levy (2021) have used field experiments to establish causal links between social media use and affective polarisation. By modifying participants' actual social media feeds and using behaviour-tracking apps, such studies achieved greater external validity with minimal researcher intervention, often in collaboration with social media platforms. While lauding such innovations, this research, instead, focuses on natural language processing (NLP) and network analysis to measure polarisation, as these computational approaches allow for large-scale analysis of real-world social media data without the artificial constraints of experimental setups. NLP can reveal partisan language patterns, while network analysis uncovers polarisation in social connections, both providing insights directly from authentic social media interactions.

I now discuss two major computational works related to the measurement of polarisation, one involving social network analysis (Barberá et al., 2015) and one related to natural language processing (Simchon et al. 2022). The former comes from a large and established body of literature that uses network analysis to study polarisation (e.g., Guerra et al., 2021; Marozzo and Bessi, 2018; Peralta et al., 2024). On the other hand, the latter comes from a more limited natural-language-processing scholarship to measure polarisation via textual data, often focusing on classifying texts by partisanship rather than divisive rhetoric (Budhiraja and Pal, 2020; Demszky et al., 2019; Nemeth, 2022; Sharber, 2020).

Barberá et al. (2015)'s *Tweeting From Left to Right* is a significant example of a network-based measurement of political polarisation in the US. The paper has estimated the ideological preferences of 3.8 million Twitter users and, studying over 150 million tweets, explored partisan groups' online communication. To do so, they have built a dataset of 3.8 million active Twitter users in the US, sampling users, who posted at least one tweet concerning at least one of 12 significant (political and non-political) events and issues between 2012 and 2014, and followed at least one of the 1,206 political accounts identified by the authors. Barberá et al. have also applied several activity, location and spam filters. Starting from the sampled users, the authors used users' following lists to compute

their ideological coordinates (ideological latent space) for those individuals who followed at least 10 of the relevant political accounts mentioned above. Their model assumes that the ideological positions of users can be inferred on the basis of observed connections among them (via their following lists). In essence, looking at the ideological positioning of the political accounts followed by the sampled users, Barberá et al. inferred users' positioning. The authors have run several robustness checks, matching users to publicly available electoral registers, and comparing their inferred versus actual ideological positioning. Having completed the users' datasets, Barberá et al. then calculated several polarisation metrics: "(a) the percentage of retweets that took place among individuals who were ideologically similar, (b) the degree of ideological homogeneity in communities of users detected in the retweet network, and (c) the average ideological extremity of the content that was spread via retweets." Their findings show that individuals, with similar ideological preferences, in the case of political issues (e.g., 2012 presidential election, 2013 government shutdown) have exchanged more information, but less regarding non-political events (e.g., 2013 Boston Marathon bombing, 2014 Super Bowl). Moreover, considering both political and non-political issues, Barberá et al. find that liberals were more likely than conservatives to engage in cross-ideological dissemination. This has led the authors to conclude that previous scholarship may have overestimated the degree of ideological segregation in social media usage.

Despite producing robust work, Barberá et al.'s findings present some limitations. First, the model considers a limited amount of political and non-political topics (12), over a limited time window 2012-2014. Such a limited list of topics may not fully reflect the dynamic nature of contemporary politics, which expands over several topics that vary in complexity (e.g., think of how the topic of immigration can be studied in its social, economic and political dimensions) and duration (e.g., immigration waves that can happen over time). Second, and relatedly, their time window does not allow researchers to study key moments in American politics, such as the Presidential elections, which may be considered a more polarised period, given the higher exposition of users to politicians and their ideologies. Third, only considering retweets does not allow Barberá et al. to investigate what people think of those political and non-political events. Specifically, it could be the case that the users are interested in sharing an event for which, however, they could hold either positive and/or negative views, which are not considered by the paper. Finally, although Barberá et al. provide a robust approach to infer users' ideological positioning, it does not allow studying users who do not

meet the following prerequisites of 10 political actors, meaning that only users who follow those actors are included in the study. As such, several users may have been excluded by the model, leading to some selection bias in the results.

More work on measuring polarisation in text over time comes from Simchon et al. (2022)'s *Troll and divide: the language of online polarization*, who have developed and evaluated a polarised language dictionary to study online Russian trolls' behaviour. To do so, they have employed a pre-coded dataset by Brady et al. (2017), who studied the diffusion of controversial political content on Twitter. In particular, they have used Brady et al.'s dataset, comprising 24,849 tweets, which were coded on whether the tweets were polarised (i.e., retweeted within one political community), or not (i.e., retweeted by a user from the opposing ideology). First, Simchon et al. have conducted a differential language analysis to compare word frequency between the two clusters (i.e., polarised and non-polarised tweets). This analysis has identified words significantly associated with polarisation. Second, the researchers have manually pruned the list of words, removing names and specific topics (e.g., Bernie, Antarctic), with a Cohen's $\kappa$ agreement rate of 0.61, to create a more generalisable dictionary. Third, word embeddings have been then used to expand the pruned dictionary, incorporating semantic fields and enhancing linguistic coverage. For example, they have expanded the word threat by the related words threats, attacks, terrorism, targets, and threatening. Finally, proper names and nonsensical additions have been manually trimmed from the expanded dictionary, resulting in a final dictionary containing 205 words. The developed dictionary has then been employed to study tweets and Reddit posts. The dictionary has accurately predicted partisan differences in COVID-19 public opinion polls, and generalised its findings to new contexts (e.g., #BlackLivesMatter) and platforms (Reddit), and studied Russian troll activity. Significantly, compared to regular American users, trolls consistently used more polarised language and increased its use over time. Finally, they have found that polarised language correlated with higher engagement, especially among politically engaged users.

However, Simchon et al. (2022)'s work does not come without limitations. First, the dictionary is limited in its scope, as it has been built solely on three topics (gun control, same-sex marriage, climate change) that have been timely in 2015. Thus, similarly to Barberá et al. (2015), this may not reflect the dynamic nature of contemporary politics. Second, and relatedly, their total number of words (205) remains very limited in number, considering the fast evolution of social media language,

which is characterised by the use of various neologisms and acronyms. Therefore, the application of the dictionary to new contexts and periods, should always be validated. Third, the dictionary produces as an output a polarisation score, which is calculated based on the count of dictionary words in the text, which is then normalised by the tweet length. However, normalising a score by tweet length assumes that the length of a tweet reflects the complexity or depth of its content, overlooking the language variability and structure across different tweets, and users. For example, a tweet may contain simple statements, while another may contain multiple clauses, citations, or references. As a result, I may expect the length of a tweet not to accurately reflect the depth or complexity of its underlying sentiment or ideological stance. Similarly, normalisation may not account for tweets' communicative purposes, which range from sharing news articles to expressing personal opinions or engaging in debates.

### 2.2.2 Wisdom of the Sub-crowds

As previously mentioned, existing polarisation measurement methods, such as network (Barberá et al., 2015) and dictionary-based approaches (Simchon et al., 2022), assume uniform perception, neglecting subjective bias. Coders may unintentionally introduce biases influenced by in-group identities and cognitive factors like motivated reasoning (Ashokkumar et al., 2020; Ecker et al., 2022; Pennycook and Rand, 2019; Pennycook et al., 2021; Taber and Lodge, 2006; Van Bavel and Pereira, 2018). The findings about online respondents point towards this idea and show that coders are able to make these associations when they label posts in response to polarisation. This has important implications for both the accuracy and the conclusions one makes about social media posts. Importantly, what researchers or pundits consider to be the ground truth in social media tone could just be a regression to the -biased- mean.

Following past work (e.g., Appel et al., 2023; Kozyreva et al., 2023; Williams et al., 2025) and extending Benoit et al. (2016)'s work, I propose a group-based "wisdom of the crowds" method to improve the measurement of polarisation, accounting for coders' partisanship. Hence, my strategy builds on previous work by Benoit et al. (2016) on crowdsourced human annotation on text analysis. The authors highlight the challenges of expert-driven data transformation in empirical social science and proposes a crowdsourcing approach for reading and interpreting text data. By involving non-experts in data collection, results comparable to those obtained by experts are

achieved more rapidly and with greater flexibility. Importantly, the transparency and reproducibility of crowdsourced datasets enhance the reliability and extendibility of research findings, shifting focus towards establishing robust methods for data collection. The approach is demonstrated to be applicable across various types of political text and languages, suggesting broader implications for expert-generated data in the social sciences. Sampling numerous workers helps reduce errors and biases inherent in individual judgments. Moreover, several statistical models, such as Bayesian scaling, are employed to aggregate information and assess worker quality, ensuring reliability in crowdsourced data generation.

A prominent example of such approach comes from Carlson and Montgomery (2017), who developed a crowdsourced pairwise comparison framework for encoding political texts, employing human coding and computational aggregation, as well as addressing concerns about coder biases and unreliability. To do so, the authors used the SentimentIt platform, a freely available cloud-based web application that simplifies interactions with Amazon Mechanical Turk (AMT) by enabling users to post jobs, certify workers, and store responses efficiently. The framework is validated using US Senate advertisements and State Department human rights reports. However, the authors notice that the coding tasks were specifically aimed at evaluating aspects of a text, which were allegedly objective; they did not ask workers to use their own judgment to code the texts. Thus, more research is needed into subjective coding. Departing from Benoit et al. (2016), new classifiers should control for coders' ideology. While approaches like "wisdom of the crowds" have been lauded for their efficiency and low cost, the impact of the heterogeneity of the coding crowd when labelling divisiveness, or polarisation, has not been accounted for sufficiently. There is evidence to argue that coders will use their in-group identity and partisanship to evaluate (and label) public discourse (e.g., Appel et al., 2023; Kozyreva et al., 2023; Williams et al., 2025). Here, I propose a new, group-based "wisdom of the crowds" strategy to improve the measurement of polarisation. To study how coding decisions are made, I investigate whether coders use their in-group identity and interest to evaluate (and label) public discourse. The results can be trivial; for concepts where the truth is in the "eye of the beholder" measurement comes with error. A Republican coder will turn a blind eye to Republican incivility, while a Democrat will be less likely to score a post as divisive when the poster is a Democrat. I believe this is not confined to partisan identity. Do Republicans and Democrats classify social media posts from the opposing party in the same way? Or, leaving

a human perspective, how would that coding differ from a generative AI? The immediate answer is that for different groups, different words might mean different things. The word "immigration" for example, is more likely to appear close to "reform" when a Democrat is speaking, and closer to "illegal" when a Republican is posting (for word embeddings, see Rodriguez and Spirling, 2021).

Recent work in such a more subjective direction has been made by Cheng et al. (2020), who compiled a large hand-coded dataset of over 13,000 COVID-19 policy announcements from more than 195 countries, providing a detailed record of government responses to the pandemic. The authors employed research assistants (RAs) to code data, controlling for their diverse, country-specific knowledge to overcome the limitations of crowdsourcing. Starting in March 2020, Cheng et al. collected 13,398 records, with each RA tracking government policy actions in at least one country, including the US, Germany, and France having multiple RAs due to complex policy coordination. The authors report the following inter-coder metrics: Percentage agreement 74%, Cohen's kappa 0.72, Krippendorff's alpha 0.71, Scott's pi: estimate (s.e.m.) 0.71. However, more research is needed to apply such an approach to coders' perspective (or, ideology) to measure concepts like polarisation over time.

And finally, it is vital to discuss Appel et al. (2023)'s approach to study the effect of partisanship on content moderation. To examine whether partisan conflict in content moderation stems from a preference gap and party promotion, the researchers embedded an experiment in a 2021 national survey of US adults, conducted by Ipsos for the Knight Foundation. They aimed to eliminate the "fact gap" by explicitly informing participants that the presented headlines were false. Each respondent viewed two false headlines: one aligned with their political party and one opposed, with the order randomised. The main outcomes measured were (i) intent to remove the headline, (ii) perception of removal as censorship, and (iii) intent to report the headline as harmful. Additional measures included perceived accuracy of the headlines. The study used randomisation at both participant and headline levels, focusing on English-speaking Democrats and Republicans. Appel et al. find a large and statistically significant difference between the content moderation preferences of Republicans and Democrats when coding headlines from opposing parties. Specifically, Democrats were more likely than Republicans to support removing false headlines (69% vs. 34%) and to report them as harmful (49% vs. 27%). Conversely, Republicans were much more likely than Democrats to view the removal of false headlines as censorship (65% vs. 29%). Consequently, my work focuses on

whether such a partisan divide may shape perception of polarisation on social media. Based on these previous accounts, I formulate the following hypothesis:

> **Hypothesis 1:** Coders are more likely to perceive posts from their in-group as less polarising and posts from out-groups as more polarising.

### 2.2.3 Using Generative AI as a Coder

While my theory provides insights into human coder behaviour, it does not offer strong theoretical explanations for AI-based tweet labelling. However, a growing body of research shows that Large Language Models (LLMs) and generative AI can perform accurate and reliable textual annotation tasks (Gilardi et al., 2023; Yang and Menczer, 2023; Ziems et al., 2023), which I now discuss in turn.

Ziems et al. (2023) show that current Large Language Models (LLMs), such as ChatGPT, perform well in data annotation tasks when prompted to directly label text out-of-the-box. In particular, regarding Misinformation, Political Stance, and Emotion classifications, LLMs achieve 70%+ accuracy when compared to ground truth, which only narrowly misses conventional accuracy standards in political sciences of 75%. However, Ziems et al. notice that ChatGPT's outputs often default to the neutral, more recognisable labels rather than use a more taxonomy-specific label. This means that coding by LLMs should always be reviewed by humans to certify accuracy. Yang and Menczer (2023), instead, use LLMs to code news outlets for their credibility (i.e., the accuracy of the information provided). Results show that LLMs, such as ChatGPT, strongly correlate with the scores given by news experts.

Gilardi et al. (2023) provides further evidence for the use of generative AI in text labelling tasks, even outperforming human annotators. In particular, they compare ChatGPT (November 2022 version)'s performance in text annotation tasks against both crowd workers, who were recruited through AMT, and trained annotators (research assistants). The authors employ a dataset of 6,183 documents, which includes tweets and news articles around content moderation. Gilardi et al. show that ChatGPT's zero-shot accuracy outperforms crowd workers in tasks, such as relevance, stance, topics, and frame detection, with an average accuracy advantage of 25%. ChatGPT also shows higher inter-coder agreement than both crowd workers and trained annotators. Their results

demonstrate the potential of LLMs to increase the accuracy and costs of text classification.

Finally, Törnberg (2024) shows that not only LLMs outperform expert human coders but also supervised classifiers at annotating political social media messages. Törnberg studies the performance of OpenAI's GPT-4[2] against human coders and supervised models in identifying political affiliations from X/Twitter messages across 11 countries. The generative AI, GPT-4, outperforms both human coders and traditional models, achieving an accuracy of 0.934 and a Krippendorf's Alpha inter-coder reliability of 0.982 in the US. Törnberg, notably, finds that LLMs excel in interpreting implicit references.

Following previous literature, I employ OpenAI's GPT-4o (Version: 2024-04-09)[3] as a benchmark of human coders' labels. Specifically, this version of GPT-4o is a multimodal AI model, which is capable of processing both text and image inputs to generate text outputs, demonstrating significantly improved accuracy and reasoning capabilities compared to its predecessors. The model training data is up to December 2023, therefore, covering the 2016-2020 US Electoral cycle under analysis. It has a large context window of up to 128,000 tokens, it excels in solving complex problems and following intricate instructions, while being 40% more likely to produce factual responses than GPT-3.5, reducing potential cases of hallucinations. This model is available through the OpenAI API[4] for paying customers. As an exploratory analysis, I will study if and the extent to which machines deviate from each partisan group. In Appendix A2, I display coding examples compared to human coders. In the next section, I discuss how AI-coding is used in this paper. Based on these accounts, I formulate the following hypothesis:

> **Hypothesis 2:** There will be significant differences between human-coded and AI-coded polarisation labels, with humans displaying greater variability influenced by their partisan identity.

[2]Törnberg (2024) employs GPT-4's version gpt-4-0314.

[3]Full commercial information retrieved via OpenAI official website. Available at: https://platform.openai.com/docs/models (Accessed on 11 October 2024).

[4]Available at: https://openai.com/api/ (Accessed on 11 October 2024).

### 2.2.4 Contributions

Based on this review, my paper seeks to fill the following gaps. First, I aim to develop supervised transformer-based classifiers to detect and measure divisive language on a 0-100 scale to fully capture its nature and magnitude. Second, I am interested in studying how coders employed to train the classifier utilise their partisanship reasoning in their polarisation judgement. This procedure allows for a direct measurement of partisan specific polarisation, which has the potential to be of great use for anyone interested in measuring the levels of divisiveness in public discourse, as well as their broader consequences. Third, I seek to employ state-of-the-art generative AI to aid me in my measurement task, as well as studying human-machine deviations. Finally, the transformer-based architecture of such classifiers will make them capable of covering longer periods, including election and non-election years, to capture full political cycles, given their powerful contextual understanding. This contribution will be explored in the next chapters of this thesis, when I apply my classifiers to the full population of tweets, and study their time and network dynamics.

## 2.3 Data and Methods

I now present the data used in this paper, as well as the research design, which is divided into three steps. First, I describe how tweets were coded via the different sets of human coders and generative AI. Second, I discuss how the levels of polarisation in the coded tweets is assessed. Finally, I discuss how the different classifiers are developed, and their performance is tested.

### 2.3.1 Data

In this paper, I develop four supervised machine learning classifiers to label tweets for containing polarising language. To do so, I utilise a large corpus of tweets posted between 2015 and 2020 in the US, developed by Friedrich, Stoehr and Formisano (2022). The tweets have been downloaded using the Twitter Academic API between October and December 2020. The tweets under analysis all contain the keyword "election"[5]. The dataset includes over 1.5 million geotagged tweets posted

[5]Friedrich, Stoehr and Formisano's full dataset also considers other keywords, such as "trump", "clinton" and "biden", as well as "brexit" for their UK case study. In this paper, I focus exclusively on the US dataset related to the keyword "elections" to centre my analysis on the electoral process rather than individual presidential candidates. This approach seeks to enhance the classifier's ability to generalise effectively to contexts beyond the US, allowing for a broader application of the classifiers and successive findings.

in the US between 2016 and 2020, allowing me to study the whole 2020 electoral cycle. This dataset is further described in Paper 2, where I investigate it over time. To train my classifiers, I randomly sample 3,000 tweets from this dataset, including 1,000 tweets from Democrats, 1,000 from Republicans, and 1,000 from the general public. Specifically, for the latter, I have initially sampled 1,500 from the general public tweets, and I have analysed them using GPT-4[6] to assess their political leaning. Only the first 1,000 tweets that have showed no clear association with either party have been included in the final sample.

### 2.3.2 Human and Generative AI labelling

The random sample of tweets (N=3,000) is coded by human coders, following previous literature (see Benoit et al., 2016; Barberá et al., 2020). I have used prolific.co to recruit American coders (aged 18 years old or above) and Qualtrics.com to collect coders' anonymous survey responses (i.e., each respondent having no distinctive character or recognition factor), and their IP addresses were deleted. Respondents are first asked basic information socio-demographic questions (gender and level of education); their ideological preferences, using a scale from one to seven, ranging from 'strong democrat' to 'strong republican', with 'independent' being at the centre of the scale. Similarly, they are asked about their interest in politics, using a scale from one to four, ranging from 'not interested' to 'very interested'. Following this, the coders are asked to code the tweets. Respondents are provided with detailed instructions on how to code the tweets for divisive rhetoric. Coders from each political affiliation have coded a random sample of 25 tweets by divisiveness, using a 3-point ordinal scale (e.g., unifying, divisive, or neither unifying nor divisive). In training the classifiers, this scale is aggregated at the respondents' ideology and dichotomised. In Appendix A1, I display the full survey, including the information and consent sheet, demographic and ideological preference questions, instructions, coding tasks, and summary statistics of coders.

To evaluate human coders' labels, as a benchmark, I employ GPT-4o to code the same texts. To do so, I use the OpenAI API[7] to run the model as follows. First, unlike previous work (Gilardi et al., 2023; Heyde et al., 2024a), I provide coding instructions and examples to the model (see

[6]GPT-4 has been prompted with the following instructions: Classify the political affiliation (either: Democrat/ Republican/ Independent/ Unknown) of the following tweet. Only provide the affiliation as an output, no explanations needed.

[7]Available at: https://openai.com/api/ (Accessed on 11 October 2024).

Appendix A1), to guarantee that the output can be compared to human annotators, having received the same instructions. Second, since previous research indicated minimal variation in predictions when prompting GPT multiple times (Heyde et al., 2024a; 2024b), I prompt the AI only once for each tweet. Third, I set the temperature to "0", limiting the creativity of the model, as I am interested in consistency and accuracy, more than innovative or unpredictable labels. Finally, to control GPT-4o's labelling task, I limit its output to a maximum of five tokens, so that the model only returns a point on the 3-point ordinal scale, as per the human coders.

### 2.3.3 Learning from the Coded Tweets

Once the coding phase is completed, I have four sets of coded tweets: Republicans, Democrats, Combined (i.e., the average score across Democrat and Republican coders) and generative AI. I use these sets of tweets to produce descriptive statistics via distribution analyses into how each partisan group and AI perceive polarisation. First, I learn on average how these groups perceive the levels of polarisation in tweets, without specifying the political affiliation of the users who posted those tweets. Second, I then study how each group perceived polarisation when the tweets are posted by a Democrat, Republican or unknown user. Appendices A2 and A3 provide an overview of my evaluation checks on human vs AI coding: qualitative visualisations of coders' dis-/agreements over the tweets' coding task, as well as a keyness analysis - computing chi-squared and p-values to understand the most important terms to predict divisiveness for each of the four subgroups. Finally, Appendix A4 provides a comparison between human and GPT-4o coding versus Simchon et al. (2022). Finally, Appendix A5 presents the results of the topic modelling analysis.

### 2.3.4 Developing Supervised-Machine Learning Classifiers

To train a supervised machine learning classifier on the human-coded tweets, I employ two variations of a pre-trained neural network, BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018) on a sequence classification exercise. Such classification is a task in which the model is trained on a set of data that includes both the input data (tweets) and the desired output (binary level of polarisation). Specifically, I employ two BERT-based models: RoBERTa[8] (Liu et al., 2019) and BERTweet[9] (Yuan et al., 2020) borrowed from the HuggingFace.co public

[8]Model available at: https://huggingface.co/FacebookAI/roberta-base (Accessed on 11 October 2024).
[9]Model available at: https://huggingface.co/vinai/bertweet-base (Accessed on 11 October 2024).

library that provides access to pre-trained models and datasets for NLP tasks. RoBERTa follows the same training procedure as BERT, which uses a masked language modelling objective (Devlin et al., 2019). However, it includes innovative training steps, such as larger batch sizes, dynamic masking patterns, and removing the next sentence prediction task. RoBERTa outperforms BERT over several classification tasks: GLUE (General Language Understanding Evaluation), RACE (Reading Comprehension from Examinations), and SQuAD (Stanford Question Answering Dataset). BERTweet also uses a masked language modelling objective (Devlin et al., 2019), as well as the pre-training procedure as RoBERTa (Liu et al., 2019). However, and most importantly, BERTweet is fine-tuned on a pre-training dataset of 850M Tweets (16B word tokens). Such fine-tuning exercises allow the model to better classify tweets, which are characterised by short texts, as well as the use of informal grammar and irregular vocabulary (e.g., abbreviations and hashtags).

Below, I select the best performing algorithm based on Accuracy, Weighted F-score and Area-under-the-curve (AUC) measures[10] on unseen testing sets. Although these models are able to uncover higher order heterogeneity, they are also sensitive to model specification and randomisation. As such, I train several classifiers with different specifications. First, I split the coded tweets into three different ratios for training, validation and testing sets (60/20/20%, 70/15/15%, and 80/10/10%). Second, setting three random seeds for each specification, I randomise the order of the tweets in the training, validation and testing sets, as well as further shuffling the tweets within their assigned batch. Third, I set the following conventional hyperparameters: early-stopping to 5 epochs; learning rate to 2e-5; and training epochs to 30. Fourth, as a robustness check, I "freeze" the entire best performing model and only train the logistic regression layer. Subsequently, I compare the performance of the frozen and full classifiers. Finally, I also compare my classifiers' performance, against a majority rule, which always predicts the most frequent class in the training set. In the following sections, I discuss the best performing classifiers, leaving the full specification (e.g., random seeds, texts' split, benchmarks, etc.) for each classifier in Appendix A6.

Following this procedure, the resulting trained classifiers take a tweet as an input, and then predict

[10] Accuracy measures the proportion of correctly classified texts out of the total number of texts. The Weighted F-score is the harmonic average of precision and recall, weighted by the support (i.e., the number of true texts) of each class, making it more suitable for datasets in which some classes are more/less frequent than others. The Area Under the Curve (AUC) is commonly used to evaluate the effectiveness of binary classification models (such as those under analysis). It represents the area under the Receiver Operating Characteristic (ROC) curve, that is, a representation of a model's true positive rate against its false positive rate at various thresholds.

the probability of the text to contain polarising language (0-100 scale). This step of the research will be conducted in Paper 2.

## 2.4 Results

This section is divided into three complementary parts. I first provide an overview of my training dataset. Second, I study the effect of coders' partisanship on their polarisation judgements. Third, I introduce my best-performing classifiers.

### 2.4.1 Descriptive Statistics of Labels

Table 1 presents a summary of the coded tweets, categorised by the party affiliation of both the coders and posters. The table provides the distribution of the coders in the dataset, having Democrat coders (1,031) compared to Republican coders (1,084). Similarly, 33% of the coded tweets have been posted by Democrats, while 33% by Republicans. The remaining tweets (33%) were not attributed to any specific party (neither Democrat nor Republican). As previously discussed, GPT-4o has been employed to code the same tweets as the human coders.

Table 1: Summary of the coded tweets by tweets' posters and coders' parties

| **Coders' partisanship** | **N Coders** | **N Tweets by Poster** | | | |
|---|---|---|---|---|---|
| | | **Democrat** | **Republican** | **Unknown** | **Overall** |
| **Democrats** | 1,031 | 1,000 (33%) | 1,000 (33%) | 1,000 (33%) | 3,000 (100%) |
| **Republicans** | 1,084 | | | | |
| **GPT-4** | N/A | | | | |

Table 2 shows the inter-coder percentage agreement[11] by coders' and posters' partisanship. Considering all tweets, the combined agreement score is 63%, with a marginal difference between Democrat and Republican coders. Democrat coders demonstrate a higher agreement rate of 70%, compared to 67% for Republican coders, indicating a higher consistency among Democrat coders in their labelling task. For tweets posted by Democrats, Republican and Democrat coders achieve the same agreement rate of 60%. The combined agreement for these tweets is slightly lower than the

[11] Percentage agreement measures the extent to which different raters or coders make the same decisions (or label classifications), and it is expressed as a percentage of total decisions.

overall rate, at 58%. For Republican-posted tweets, Democrat coders exhibit a strong agreement of 64% compared to 59% for Republicans. The combined agreement for Republican posters is 60%, slightly lower than the overall average rate. For tweets where the poster's political preference is unknown, Democrats maintain a higher agreement rate of 64%, while Republicans have a lower rate of 61%. The combined agreement for these tweets is 61%, once more lower than the general average. To summarise, Democrat coders consistently show marginally higher agreement rates across all categories of tweet posters. However, overall agreement among coders fluctuates, depending on the poster's party affiliation.

Table 2: Summary of intra-coder agreement by party ideology and tweets' posters

| **Posters** | **Coders** | | |
|---|---|---|---|
| | **Democrats** | **Republicans** | **Combined** |
| **All tweets** | 70% | 67% | 63% |
| **Democrat** | 60% | 60% | 58% |
| **Republican** | 64% | 59% | 60% |
| **Unknown** | 64% | 61% | 61% |

Table 3 presents the percentage agreement between different sets of coders: Democrats, Republicans, Combined and GPT-4o. The results illustrate how closely each group agrees with one another in their coding of tweets. The agreement between Democrat and Republican coders is 80%, indicating high alignment between these two groups. As expected, both Democrat and Republican coders show strong alignment with the Combined coders' scores, with 89% and 90%, respectively. In contrast, Democrat coders show a much lower level of agreement with GPT-4o (47%). Similarly, Republicans also show low agreement with GPT-4o (46%), indicating a relatively weaker alignment between the two. Also the Combined scores has a weak agreement with GPT-4o (44%). Therefore, while there is a strong agreement between the human coders within each political group and their Combined scores, there is a considerable disconnect between these coders and GPT-4o. The lower agreement with GPT-4o highlights potential differences in how machine-generated classifications align with human coders' interpretations.

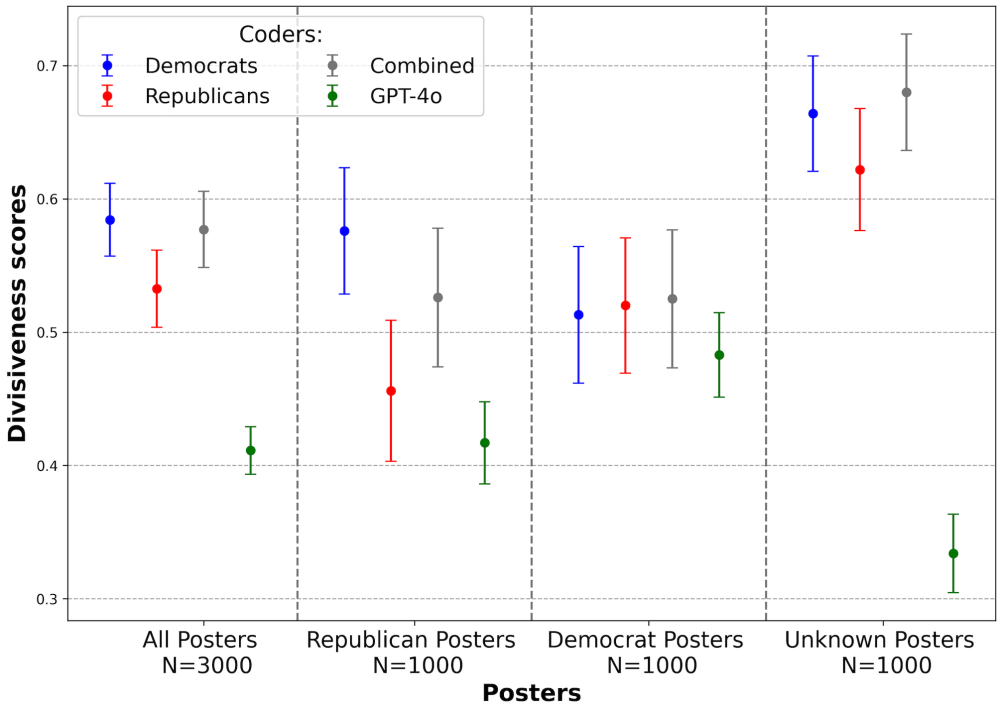
Table 3: Summary of percentage agreement between inter-coders

| - | **Democrats** | **Republicans** | **Combined** | **GPT-4o** |
|---|---|---|---|---|
| **Democrats** | - | 80% | 89% | 47% |
| **Republicans** | 80% | - | 90% | 46% |
| **Combined** | 89% | 90% | - | 44% |
| **GPT-4o** | 47% | 46% | 44% | - |

### 2.4.2 Coders' Partisanship

Having discussed the composition of the training dataset, Figure 1 compares divisiveness scores across different coders and posters. It explores how the partisanship of coders—Democrats, Republicans, and their combined coding— affects their assessment of divisiveness in tweets, also including results from the AI-coding. The y-axis represents the divisiveness score, while the x-axis groups different poster categories, such as "All Posters" (N=3,000), "Republican Posters" (N=1,000), "Democrat Posters" (N=1,000), and "Unknown Posters" (N=1,000). Each group is presented with the mean divisiveness score along with 95% confidence intervals (CI).

Figure 1: Comparison of Divisiveness Scores by Posters and Coders




*Notes: Coders' Partisanship and Tweets' Posters considered. Mean and 95% CI provided. Democrats are represented by blue markers, Republicans by red, GPT-4 by green, combined human-coders by grey.*

The plot reveals several insights into how coders' partisanship affects their assessment of divisiveness in tweets, which I discuss in turn.

Across all tweets, Democrat coders assign higher divisiveness scores than Republican, Combined, and GPT-4o coders. However, this difference is not statistically significant when compared to Republican and Combined coders. In contrast, Democrats rate tweets as significantly more divisive than GPT-4o. For Republican posters, a similar trend emerges: Democrats perceive higher levels of polarisation compared to Republicans, while GPT-4o assigns the lowest divisiveness scores among all coders. In this group of posters, the Democrats' ratings are significantly higher than Republicans' and GPT-4o' codings. For Democrat posters, Combined coders assign the highest levels of polarisation, followed by the Republicans who in turn assign higher divisiveness scores than Democrats. GPT-4o again assigns the lowest divisiveness scores. However, all of these differences are not statistically significant. Finally, considering unknown posters, the pattern remains consistent. Human coders perceive higher levels of polarisation than GPT-4o. Specifically, Combined coders

assign higher divisiveness scores than Democrats, which precede Republicans and GPT-4o. The difference between the human coders is not statistically significant, however, it is only significant when compared to GPT-4o's coding.

These results are partially aligned with my expectations. Specifically, the human coding results align with Hypothesis 1, suggesting that coders exhibit partisan bias in their assessments. Democrats consistently rate Republican posters as more divisive than Republicans do, while Republicans, in turn, assign higher divisiveness scores to Democrat posters than Democrats themselves. However, the differences between partisan coders are not always statistically significant, suggesting that while bias may be present, its impact varies across different tweet posters. Notably, when examining tweets posted by Republicans, where Democrat coders perceived significantly higher divisiveness than their Republican counterparts. Moreover, the findings lend support to Hypothesis 2, which posits a divergence between human and AI coders. Across all groups, GPT-4o systematically assigns lower divisiveness scores compared to human coders, with the difference being statistically significant in several cases. This reinforces the idea that human coders, influenced by their ideological perspectives, perceive greater polarisation in online discourse, whereas AI coding remains more neutral or conservative in its assessment of divisiveness.

### 2.4.3 Classifiers' Performance

Having analysed my training dataset, I train four supervised machine learning classifiers, and I report their performance across the different transformer models and specifications used. Before doing so, I dichotomise the polarisation scores, and I apply standard textual pre-processing steps, such as removing URLs, numbers, and potential duplicates to eliminate noise and train-test contamination. In Table 4, I report the best performing classifiers, which have been BERTweet for Democrat and Combined coders, and roBERTa for Republicans and GPT-4o coding. Appendix A6 contains a descriptive analysis of each classifier's performance metrics using all configurations (e.g., coding dataset, model, seed, training/validation/testing split ratios).

The performance metrics of various classifiers reveal distinct patterns in their accuracy and effectiveness in labelling tweets for divisiveness. The top-performing model, trained on Combined coders, outperforms all others, achieving an accuracy and a weighted F-score of 91%, with an AUC of 0.84. The second-best model, trained on Republican coders, follows closely with an accuracy and

a weighted F-score of 88%, and a higher AUC of 0.87. The Democrat-based model also achieves an accuracy and a weighted F-score of 88%, but has a lower AUC of 0.81. Lastly, the GPT-4o-based model attains 86% accuracy and a weighted F-score, with an AUC of 0.86.

These findings highlight the advantages of integrating multiple perspectives in the classification process, helping to mitigate partisanship biases. The superior performance of the Combined coders model suggests that the biases of Republican and Democrat coders may counterbalance each other, making the training task easier and leading to a more balanced and objective assessment of divisiveness in tweets. Notably, while the GPT-4o-based model performs well, it does not surpass human coders in classification effectiveness. This underscores the continued value of human judgment in nuanced tasks, even as AI demonstrates strong natural language processing capabilities. Some possible reasons AI underperforms compared to human coders is its built-in safeguards, which may make it overly cautious in labelling divisive content. Additionally, AI models are trained on vast amounts of political data, potentially making them more informed than the average American, but also more prone to overanalysing context in ways that diverge from human judgment. While AI is highly consistent in its coding decisions, this consistency may come at the cost of missing contextual nuances that human coders can better interpret.

Table 4: Summary Training Performances by Coders and Specification

| Coders | Model | Seed | Train/Val/Test Split | Accuracy | Weighted F-score | AUC |
|---|---|---|---|---|---|---|
| Combined | roBERTa | 552 | 80/10/10% | 91% | 91% | 0.84 |
| Democrat | | 833 | | 88% | 88% | 0.81 |
| Republican | BERTweet | 230 | | 88% | 88% | 0.87 |
| GPT-4o | | 912 | 70/15/15% | 86% | 86% | 0.86 |

*Notes: The table shows the descriptive statistics of the best-performing algorithms on the testing sets for each set of coders. I present the model, random seed and split used. I also provide four performance metrics: Accuracy, Weighted F-score, and Area-under-the-curve (AUC) for each class in the model. Appendix A6 contains a descriptive analysis of each classifier's performance metrics using all configurations (e.g., coding dataset, model, seed, training/validation/testing split ratios).*

## 2.5 Discussion

The goal of this paper has been to improve the measurement of polarisation in public discourse. To do so, I have developed four powerful supervised machine learning classifiers to annotate text for polarising language, along a continuous scale, accounting for coders' party affiliation, as well as employing generative AI as a benchmark.

Analysing the coding datasets, I find that Democrat coders rate tweets as more divisive than Republicans, Combined coders, and GPT-4o, this difference is not always statistically significant—except in comparison to GPT-4o, where Democrats perceive significantly more polarisation. A similar trend appears for Republican posters, where Democrats assign notably higher divisiveness scores than both Republican coders and GPT-4o. However, when evaluating Democrat posters, the Combined coders perceive the highest levels of divisiveness, followed by Republicans, Democrats, and finally GPT-4o, although these differences do not reach statistical significance. The same general pattern holds for tweets from unknown posters, with human coders consistently rating divisiveness higher than GPT-4o. Among human coders, Combined scores are the highest, followed by Democrats and Republicans, but these differences remain statistically insignificant. In contrast, for these posters, the gap between human coders and GPT-4o is significant.

The results provide partial support for Hypothesis 1, suggesting that partisan biases influence human coding. Democrats rate Republican tweets as more divisive than Republicans do, and vice versa, though the magnitude of this difference varies across groups. Notably, this bias is statistically significant when Democrats evaluate Republican posters, reinforcing the idea that partisan identity affects perceptions of polarisation. However, the fact that some differences do not reach significance suggests that partisan bias is not always strong enough to create major divides in coding outcomes.

Meanwhile, Hypothesis 2, which posits a systematic difference between human and AI coders, is strongly supported. In every group, GPT-4o assigns the lowest divisiveness scores, diverging significantly from human ratings in multiple cases. This suggests that human coders, influenced by their ideological perspectives, perceive greater polarisation, while AI remains more neutral or conservative in its assessments of divisiveness.

Turning to classifier performance, the best-performing model is trained on Combined coders,

achieving an accuracy and a weighted F-score of 91%, with an AUC of 0.84. The Republican-trained model follows closely, achieving 88% accuracy and weighted F-score, with an AUC of 0.87, while the Democrat-trained model also reaches 88% accuracy and weighted F-score, but has a slightly lower AUC of 0.81. Finally, the GPT-4o-trained model performs the weakest, with 86% accuracy and an AUC of 0.86.

These findings underscore the value of integrating multiple perspectives. The superior performance of the Combined coders model suggests that partisan biases may counterbalance each other, resulting in a more balanced assessment of divisiveness. While GPT-4o demonstrates strong consistency, it does not outperform human coders when training algorithms in classification accuracy, reinforcing the importance of human judgment in nuanced tasks. One possible reason for GPT-4o's lower divisiveness scores is its built-in safeguards for political neutrality, which may make it more cautious in labelling content as divisive. Additionally, AI models are trained on vast political datasets, potentially making them more context-aware than the average human coder but also more prone to overanalysing content in ways that diverge from human judgment.

These results provide several contributions to the computational political science literature. First, I introduce innovative machine learning classifiers that enhance the measurement of polarising language in social media content, which can be used in different contexts, languages and text types given its transformer-based models. Second, I provide a critical evaluation of the factors influencing coding decisions, particularly the impact of coders' partisan affiliation, which has not been attempted before in the context of polarisation, contributing to pre-existing work (e.g., Appel et al., 2023; Kozyreva et al., 2023). Third, I show how human coding compared to generative AI. The integration of algorithm development with a thorough analysis of human biases marks a significant advancement in the field, offering valuable tools for researchers and practitioners seeking to identify and measure polarisation in public discourse.

While the findings offer important insights, some limitations should be acknowledged. First, my analysis does not investigate the demographic characteristics of the human coders. In other words, I do not examine whether gender, age, education, or occupation influence the coding process. Investigating all the socio-economic factors shaping perceptions of polarisation is beyond the scope of this study. However, analysing respondents' ideological preferences should be considered

essential, as political beliefs may directly impact how individuals perceive and label polarisation. In Appendix A7, I evaluate socio-tropic versus ego-tropic considerations in line with previous literature by Margalit (2013), and more recently by Schaffer and Spilker (2019). Finally, in Appendix A8, I then assess Williams et al. (2025)'s hypothesis on the effect of gender on subjectivity. In both cases, I find that my results are confirmed and coders' partisanship is the most important aspect in predicting coders' perception of polarisation, outweighing the influence of other demographic (i.e., gender) or socio-economic (i.e., socio vs ego-tropic considerations) variables.

Another aspect to discuss is the possibility that my results may be driven by the highly polarising political context, party system and the English language used in the US. Specifically, my measurement may not apply to multiparty systems, emerging democracies, and/or different languages. In other contexts the driver of polarisation may be different: in Europe, for example, party systems are more fragmented and citizens exhibit less partisan loyalty (Huddy et al., 2018; 2020; Kriesi et al., 2006; 2012), or in Latin America, where polarisation often centres around a strong personalisation of the party system with charismatic and populist leaders and movements (Areal, 2022; Torcal and Carty, 2023; Torcal and Harteveld, 2024). Similarly, in new democracies, institutional trust and party systems may still be weak, polarisation could be driven more by ethnic, religious, or regional cleavages rather than traditional partisan identities (Reiljan et al., 2024; Torcal, 2023; Torcal and Harteveld, 2024). Finally, language may differ from other anglophone countries (e.g., the UK, Australia).

However, having trained a context-based machine learning classifier, the encoded layers provided by BERTweet and roBERTa should be able to capture different political domains and linguistic variations (for example, between British and American English language). Such algorithms can be updated using other models, such as mBERT[12] and XLM-roBERTa[13], which function well in multilingual settings. Similarly, to train the classifiers, I have utilised tweets, which are dense forms of text (given their 280-character limits), therefore I expect that this classifier can be potentially used for longer texts, for example, press releases and Parliamentary speeches. Nevertheless, I urge caution to always validate my classifiers' predicted scores both via quantitative (e.g., checking for the distribution of the coded categories) and qualitative (i.e., analysing a small sample of the full

[12]Model available at: https://huggingface.co/google-bert/bert-base-multilingual-cased (Accessed on 23 October 2024).

[13]Model available at: https://huggingface.co/FacebookAI/xlm-roberta-base (Accessed on 23 October 2024).

population of the ML-coded texts to ensure the correct coding of unseen data).

In the next chapters, I will employ these four polarisation classifiers to code and analyse the full dataset of US tweets. Paper 2 will study polarisation over time and topics, and, more importantly, it will investigate whether polarisation comes from political actors and/or the public and/or news media actors. Paper 3 will focus on the network structure of polarisation, exploring how politicians across the political spectrum position themselves. The combination of these three chapters seeks to provide a comprehensive review of the dynamics of polarisation in US Politics on Twitter (2015-2020).

# 3 The Temporal Dynamics of Online Polarisation in the US

**Abstract**

Affective polarisation, the deep emotional divide between political groups, poses a significant threat to contemporary American democracy. However, the precise mechanisms by which polarisation spreads and the direction of its influence between politicians, news media, and the public remain unclear. While some scholars contend that polarising language originated among the politicians to the public, others argue that it instead emanates from the public sphere and influences politicians, while news media act as catalysts. To adjudicate among these competing theories, I analyse a dataset of tweets posted in the US between 2015 and 2020 by US politicians, news outlets, and the public. Polarisation levels are measured using supervised machine learning algorithms developed by human coders, accounting for their political affiliation, and generative AI. Overall, the results reveal competing findings when accounting for who posts and reads the social media content. Vector-autoregressive analyses show the reciprocal and evolving nature of online polarisation, finding that both parties influence each other's rhetoric over time. While the media acts as a catalyst for polarisation, politicians lead and follow the public in shaping discourse, with cross-polarisation interactions contributing to an ongoing cycle of political division. This study provides new insights into the dynamics of polarisation, with implications for democratic discourse and civic engagement in the US and beyond.

**Keywords:** Affective Polarisation, Machine-Learning, Time-series, US Politics, Twitter.

## 3.1 Introduction

As seen in the previous chapters of this thesis, polarisation on social media has emerged as a critical concern in contemporary political discourse, which is often fuelled by highly emotionally charged and ideologically extreme language. While previous studies have examined the role of social media in fostering echo-chambers and reinforcing pre-existing biases (Bail, 2021; Barberá, 2020; Hobolt et al., 2024; Sunstein, 2018), less is known about who initiates such a use of polarising language and under what circumstances it escalates across politicians, news media and public users. Moreover, previous work has showed that social media platforms provide a fertile ground for polarisation because they allow like-minded individuals to congregate and exchange views in

isolated environments (Pariser, 2011; Sunstein, 2018). Politicians and political elites, in particular, may initiate or amplify negative discourse to mobilise their base (Bail, 2021), while public users, often from homophilic groups, might respond to these cues or initiate their own (Barberá et al., 2019; Theocharis et al., 2020). A such, recent scholarship has highlighted the importance of studying the temporal dynamics of polarisation, as well as investigating the actors that shape such divisive rhetoric (Harteveld, 2021; Iyengar and Wagner, 2024; Röllicke, 2023; Stavrakakis, 2018; Westphal, 2024). Specifically, more established research points to the changing nature of polarisation during salient political events such as elections, scandals, but also policy changes (Mouffe, 2000; Goi, 2005; Westphal, 2024). Regarding the actors, it is important to investigate the difference between 'horizontal' and 'vertical' polarisation (Harteveld, 2021; Harteveld et al., 2022; Röllicke, 2023; Westphal, 2024: 10). The former refers to discussions among citizens (Harteveld, 2021; Harteveld et al., 2022). In contrast, the latter refers to citizens' opinions towards parties and political elites, as a bottom-up phenomenon (Harteveld, 2021; Harteveld et al., 2022; Westphal, 2024), or it identifies a top-down phenomenon whereby parties and political elites utilise divisive language, or behaviour more broadly, towards citizens (Stavrakakis, 2018; Westphal, 2024). In this public arena, previous work has identified the media to act as an amplifier between the public and the politicians (Wlezien, 2024; Wlezien and Soroka, 2024). Thus, the temporal dynamics of these exchanges and initiating actors—whether politicians or public users more often initiate such language and which topics or events trigger heightened polarisation—remain elusive. These gaps in the literature motivate the following research question:

> *Who initiates polarising language on social media, and under what conditions do these interactions escalate polarisation?*

To answer this question, I apply the four supervised machine learning algorithms in Paper 1 to measure polarisation in a comprehensive dataset of tweets posted between 2015 and 2020 by US politicians, news media, and the public discourse on Twitter. Each classifier reflects the coding of Democrat, Republican and Combined human coders, as well as the one trained on GPT-4o coding. By employing these classifiers, I study how each coder groups perceive the levels of polarisation for each tweet in the dataset.

Following the previous empirical literature (e.g., Barberá et al., 2019; Gilardi et al., 2021; Widmann,

2022; Wlezien and Soroka, 2023; Wlezien, 2024), I conduct time-series analyses to study 'who leads' and 'who follows' in the spread of polarisation. Specifically, I run vector autoregressive (VAR) models, using the impulse-response-functions (IRFs) to study the temporal variation of polarisation at a highly granular level. These tools allow me to understand who, between different actors (political actors, public, and media), is responsible for spreading polarisation.

Time-series analyses of observational data show that Republican politicians exhibited the highest levels of polarisation from early 2015 to late 2016, after which Democrats led in polarisation from late 2016 to early 2019. From 2019 onward, the public surpassed politicians in polarisation. Key events, such as the occupation of the Malheur National Wildlife Refuge, Trump's election, mass shootings, impeachment proceedings, and the Capitol Hill riot, were associated with notable spikes in polarisation across all groups. Twitter's platform updates in April 2019 may have also played a significant role in influencing public polarisation. Finally, media posts generally exhibited the lowest levels of polarisation among all groups. The VAR analyses confirm the reciprocal and evolving nature of online polarisation, showing that both parties influence each other's rhetoric over time. While the media acts as a catalyst for polarisation, politicians lead and follow the public in shaping discourse, with cross-polarisation interactions contributing to an ongoing cycle of political division. These findings highlight the complex and dynamic nature of online polarisation and emphasize the importance of considering temporal factors in future research.

## 3.2 Who Leads, Who Follows in Spreading Polarisation?

To comprehensively study who initiates polarising language on social media, and the conditions under which these interactions escalate polarisation, I divide my theoretical framework into three-complementary steps. First, I study the interactions between politicians and the public. Second, I examine the cross-cutting interactions between opposing politicians. Third, I shed light on the role of news media in public discourse.

Political science research consistently shows that politicians often follow the preferences of their supporters rather than setting the tone themselves (Barberá et al., 2019; Gerber and Morton, 1998; Holbrook and McClurg, 2005; Egan, 2013; Kastellec et al., 2015; Neundorf and Adams, 2018). This tendency to follow is particularly evident in high-stakes moments like primaries and general

elections, where alignment with voter sentiment is crucial for securing support. Politicians actively shape their messaging to resonate with their base, carefully mirroring public preferences rather than challenging them, a strategy that reflects a calculated alignment with their supporters' views (Bawn et al., 2012; Holbrook and McClurg, 2005; Gerber and Morton, 1998). Moreover, it is a political strategy that extends to the use of emotional rhetoric to please the electorate (Crabtree et al., 2020; Kosmidis et al., 2019; Müller, 2022; Valentim and Widmann, 2021; Widmann, 2022). This type of communication can help some political actors create specific narratives, which can, in turn, lead to societal division (e.g., Lasco and Curato, 2019), mobilise support, and influence attitudes and public opinion (Kühne et al., 2011; Kühne and Schemer, 2015; De Vries et al., 2021; Yam et al., 2020; Widmann, 2022). Therefore, such a dynamic underscores a broader view of political communication as an interactive process shaped by both demand-side factors; for example the preferences and emotions of voters, and supply-side factors, such as the strategic choices made by politicians (Barberá et al., 2019). Instead of independently setting the agenda, politicians are highly attuned to the sentiments of their supporters, particularly in the digital age where social media amplifies voter engagement. Online platforms, such as Twitter, provide real-time feedback loops that politicians to gauge public sentiment and adjust their messaging accordingly. In such environments, political actors are likely to echo the emotions, frames, and narratives that resonate with their voter base, thereby reinforcing their alignment with the electorate (Hawkins et al., 2020). This behaviour not only helps politicians maintain their support but also reflects a broader trend of emotional and responsive communication in modern political campaigns.

Research on social media usage also shows that politicians primarily use platforms to broadcast their views but rarely engage directly with the public, partly due to the risks of negative responses or losing control of their message (Stromer-Galley, 2000; Ward and Lusoli, 2005; Tromble, 2018). While early studies characterised this interaction as unidirectional, more recent work highlights the potential for reciprocal exchanges between politicians and citizens. According to early supply-side theories, political actors preferred not to engage with popular political discussions due to time constraints, low levels of technological skills, or fear of losing control of the messages (Stromer-Galley, 2000; Ward and Lusoli, 2005). With the advent of social media, public users enjoy some anonymity, which allows them to express their ideas online, often harassing politicians in various ways (Theocharis et al., 2020). Tromble (2018), for the supply-side, has studied bi-directional

Twitter discussions (e.g., discussions from politicians to the public, and vice versa), showing that politicians avoid discussing with users to avoid the risk of negative reciprocity. In particular, he studied the role of users' tweets to and from members of the legislature in the Netherlands, the UK, and the US in 2013. Investigating a large dataset of tweets and replies, as well as manual annotation for tweets' sentiment (e.g., positive, negative, neutral), Tromble conducted several regression analyses, showing that users' opinions are important in determining whether and how political actors respond either through replies or future posts. This influence is substantial, though under-explored in terms of topic and initiation—key factors that may drive polarisation. Expanding on Scholl et al. (2021), I seek to study the initiation and evolution of polarising language between the public and political actors on social media. Specifically, I investigate the conditions under which tweets from politicians lead or follow the public in terms of levels of polarisation over time, testing the hypothesis that public sentiment leads and politicians respond accordingly.

> **Hypothesis 1:** Politicians are likely to adopt and amplify the polarising language of public users rather than initiate it.

The role of the media in discussions between politicians and the general public has been studied by Wlezien (2024), and Wlezien and Soroka (2024). Although these two papers do not focus on levels of polarisation, I take their account to study the direction of the relationship between politicians, general public and media over time. I now discuss them in turn. First, Wlezien (2024) examines the relationship between news coverage and public opinion, challenging the traditional assumption that media directly shape public perceptions and preferences. Instead, the author argues for a bidirectional flow of influence, where public opinion can also drive the nature of media coverage. Using three case studies (e.g., economic perceptions, candidate support, and policy preferences), the analysis employs automated content analysis and vector autoregression to explore these dynamics. The findings reveal that public opinion often influences news coverage more significantly than news affects public opinion. This suggests that public sentiment actively shapes media content, and that this influence should be studied empirically rather than presumed one-sided. Second, Wlezien and Soroka (2024) add public policy to the equation. As such, they explore the relationship between mass media, public opinion, and government policy, challenging the common view that media have a major, direct impact on democratic politics. They find that media content is not a driving force but

rather a reflection of public preferences and political actions. By examining US news coverage and its interactions with budgetary policy and public opinion in areas (e.g., defence, welfare, and health care), Wlezien and Soroka (ibidem) find that the media mirror these political and public factors rather than shaping them. In other words, they suggest that the media often amplify existing public sentiments and policy trends rather than creating them.

Based on these accounts, I argue that the media acts a catalyst in public discourse, suggesting that media primarily reflect and amplify existing public and political actions. Bringing such a discussion back to polarisation, my argument implies that media likely amplify pre-existing divides in public sentiment rather than creating them. When the general public and politicians use polarising language, media coverage tends to mirror and reinforce these shocks, potentially deepening divisions by circulating content that reflects heightened public sentiment. This feedback loop could exacerbate polarisation, as media coverage continuously mirrors polarised views, making the divides more visible and entrenched. In other words, this framework reframes media influence as primarily responsive, where the public and politicians' decisions drive much of the levels of polarisation. Thus, my expectations are that changes in divisive rhetoric by politicians and the public predict shifts in media coverage. Conversely, while the media can influence public discourse, this effect is weaker. Taking together these two aspects, I formulate the following hypothesis.

> **Hypothesis 2:** The media acts as an amplifier between politicians and the public, reflecting and magnifying shocks in the general public and political actions rather than initiating them.

Having defined how each user type may behave in the share of polarising language, the next analytical step is to define how partisanship of such a user plays a role. Cross-cutting interactions stand in contrast to ideological segregation by encouraging individuals to engage with users from different political groups, which can help reduce polarisation (Holbert et al., 2010; Bright, 2018). As such, interactions between users from different partisanships on social media introduce opportunities for users to encounter opposing views, creating pathways for either reducing or, in some cases, intensifying polarisation, contingent on the dynamics of each interaction (Barberá et al., 2015; de Zúñiga and Valenzuela, 2011). This moderating effect is given by the nature of social media, which promotes these interactions, enabling users to connect with others regardless of their political

interests or followers (Barberá et al., 2015; de Zúñiga and Valenzuela, 2011; Duggan and Smith, 2016; Newman et al., 2017). Furthermore, the desire for diverse ideas and information may push users to interact with various groups (Knobloch-Westerwick and Kleinman, 2012; Dubois and Blank, 2018). As such, cross-cutting interactions can lead to more moderate political debates and help bridge the ideological gap between different groups (Hong and Kim, 2016; Bright, 2018).

However, discussions involving moderate users from different parties are less frequent, as they often arise from variations within groups and interactions across groups (Bright et al., 2018). Bright et al.'s analysis of Twitter activities from 115 political groups in 26 countries shows that extreme groups are less likely to engage with one another compared to centrist parties, suggesting that polarisation may diminish among centrists while increasing at the extremes. For those on the extremes, cross-cutting interactions can reinforce existing beliefs and deepen polarisation, while centrists may experience reduced polarisation (Conover et al., 2011; 2012; Levendusky, 2013). Therefore, I seek to explore when and how these cross-party interactions happen. Understanding how belonging to specific political groups influences behaviour and both short- and long-term effects is crucial. In particular, it is important to investigate whether and how some political topics lead to increased polarisation among like-minded individuals or opponents.

To explain online polarisation, many scholars argue that social media platforms create echo chambers that amplify the ideas of extremists while isolating moderate users (Barberá, 2020; Sunstein, 2018; Bail, 2021). Bail (2021) complements this view with an analysis of Twitter posts and engagement, as well as survey data, finding that extremists are often pushed towards more radical positions due to the positive feedback they receive from likes and new followers. He notes that extremists might hesitate to show support for moderate views due to the fear of backlash. The desire for more engagement (such as likes and shares) can further fuel polarisation. Similarly, Sunstein (2018) claims that social media platforms make it easier for like-minded individuals to connect, increasing exposure to hyper-partisan content and reinforcing existing views while isolating users from opposing perspectives. These insular discussions can create echo chambers (Pariser, 2011; Benkler et al., 2018; Barberá, 2020), which contribute to party-line segregation, escalate political discourse, and foster public distrust in institutions, politicians, and the media. Furthermore, Hobolt et al. (2024) develop and test the argument that the political homogeneity of people's social environment shapes polarisation. Using an innovative, large-scale experiment in the UK, they examine how polarisation is influenced

by partisan group homogeneity. Empirically, Hobolt et al. recruit nationally representative partisans and assign them to discuss a salient policy issue, either with like-minded partisans (an echo chamber) or in a mixed-partisan group. This allows the authors to examine how group composition affects polarisation. Hobolt et al. find that partisan echo chambers lead to higher levels of both polarisation on policy and affective polarisation compared to discussions in mixed groups.

However, more evidence is needed to understand how these groups form and persist, especially in relation to different discussion topics and over time (Harteveld, 2021; Iyengar and Wagner, 2024; Röllicke, 2023; Stavrakakis, 2018; Westphal, 2024). Significantly, Bail (2018) suggests researching how polarisation continues after users from a group engage with others over time, stressing the importance of finding less biased methods to study polarisation, which is often overstated by heavy social media users. Moreover, it is significant to investigate the exchanges between users belonging to either politics, news media and the general public, and study who influences whom (e.g., Harteveld, 2021; Wlezien, 2024; Wlezien and Soroka, 2024; Westphal, 2024).

In this paper, I expect cross-cutting interactions to happen when users from different partisan groups discuss elections on the same day. As such, it is likely that these partisan users adopt more extreme (or polarising) positions to gain attention and support on social media platforms over time (Bail, 2018; Lin et al., 2023). However, I also expect some potential depolarising effects if partisans do engage in cross-partisan discussions, particularly on issues where there is some common ground or cross-party consensus (de Jong, 2024). For news media, it is likely that they will contribute to echo chambers by catering content to partisan audiences, reinforcing existing beliefs rather than challenging them (Lin et al., 2023).

Based on previous work, I argue that cross-interactions between users of differing partisan groups may influence polarisation levels, creating circumstances where ideological distance could either widen or narrow, depending on the participants involved and topics discussed.

> **Hypothesis 3:** Cross-interactions among users belonging to different partisanships may increase levels of polarisation.

Overall, my theoretical framework can be summarised as follows: Politicians are likely to follow the public rather than lead, while the news media users will have a catalyst role. Additionally, cross-

cutting interactions between users with different partisanships can occur, increasing polarisation among them overall, or decreasing on common/shared political grounds. Such a framework is assessed via observational and time-series analysis as described below.

## 3.3 Data

In this section, I present the dataset under analysis, as well as explaining how each tweet is coded by levels of polarisation (via the four algorithms developed in Paper 1) and topic, using unsupervised topic modelling (via BERTopic; Grootendorst, 2022).

### 3.3.1 Data Collection

As explained in Paper 1, this thesis employs the Twitter dataset from Friedrichs, Stoehr, and Formisano (2022), which studies public discourse in US politics between 2015 and 2020. The dataset, collected via the Twitter Academic API, includes all tweets containing the word "election" posted in the US between January 1, 2015, and December 31, 2020. Friedrich, Stoehr and Formisano's full dataset also considers other keywords, such as "trump", "clinton" and "biden", as well as "brexit" for their UK case study. In this paper, however, I focus exclusively on the US dataset related to the keyword "elections" to centre my analysis on the electoral process rather than individual presidential candidates. This approach seeks to enhance the classifier's ability to generalise effectively to contexts beyond the US, allowing for a broader application of the classifiers and successive findings. Notably, Friedrichs, Stoehr, and Formisano (2022) manually categorised the users who posted the tweets into three groups: political actors, news media, and the public. Moreover, political actors and news media have been further categorised by their political affiliation (Democrats and Republicans). For the general public, although previous work by Barberá (2015)'s Bayesian Ideal Point Estimation Using Twitter Data, successfully estimated users' ideological estimation, such a tool is no longer available at the time of writing due to Twitter's 2023 Terms and Conditions[14]. Table 1 provides a summary of the total number of tweets by user type.

[14]Twitter announcement made via a post on its platform in February 2023. Available at: https://x.com/XDevelopers/status/1621026986784337922 (Accessed on 17 October 2024).

Table 1: Summary of the Friedrichs et al. (2022)'s Twitter dataset subset under analysis

| User type | N tweets |
|---|---|
| Dem Political Actors | 111,129 |
| Rep Political Actors | 87,402 |
| Public Users | 674,732 |
| Dem Media | 81,810 |
| Rep Media | 92,333 |

### 3.3.2 Coding Tweets by Levels of Polarisation

To code these tweets for levels of polarisation, I employ the four algorithms that I have developed in Paper 1 (Democrats, Republicans and Combined coders, and GPT). The algorithms take the tweets' texts as an input, and return a continuous score, ranging from 0 to 1, to indicate the levels of polarisation. Section 5 displays initial descriptive statistics into the coded datasets using all coders, and accounting for all posters (Republican, Democrat, Public users, Dem Media, and Rep Media).

### 3.3.3 Conducting Topic Modelling

The next step is to determine the topic discussed in each tweet, beyond simply containing the keyword under analysis (i.e., election). To do so, I employ an unsupervised topic modelling classifier, BERTopic[15] (Grootendorst, 2022). BERTopic uses transformer models and class-based Term Frequency-Inverse Document Frequency (TF-IDF)[16] to create dense topic clusters from textual data. The process involves document embedding using pre-trained transformer models, dimensionality reduction with UMAP (i.e., Uniform Manifold Approximation and Projection), and clustering with HDBSCAN (i.e., Hierarchical Density-Based Spatial Clustering of Applications with Noise). Unlike traditional methods, BERTopic treats all documents within a cluster as a single document and applies clustering to identify relevant words. A key advantage of BERTopic is its ability to automatically determine the optimal number of topics across a textual data corpus, eliminating the need for manual specification. Such flexibility makes the algorithm effective for diverse text data analysis and producing human-interpretable results. As BERTopic takes a list of texts as input

[15]Model available at: https://maartengr.github.io/BERTopic/index.html (Accessed on 4 November 2024).

[16]TF-IDF is a numerical coefficient that measures the significance of a word within a document relative to a collection of documents, meaning that it can the words that are most relevant and distinctive to a particular document or set of documents.

and returns a predicted topic with a probability score, as well as additional information[17], scholars can fine-tune the topic representations to improve interpretability and coherence. In this paper, I have set the embedding[18] and representation[19] models. Moreover, I have first run the model on the full list of texts, and then applied the topic reduction function to reduce the number of topics to a maximum of 30 for better representation and coherence. Appendix A10 contains full information on the topics contained in each dataset of tweets. These topics will be further analysed in the next stages of this thesis.

## 3.4 Methods

To investigate who initiates the use of polarising rhetoric more granularly among politicians, news media and general public, I run vector-autoregressive (VAR) models via impulse response functions (IRFs) on stationary time-series. Below, I discuss the pre-processing and stationarity checks of the time-series, then I move to the VAR models, together with their validation checks.

### 3.4.1 Pre-processing Time-Series

Before engaging in my analytical work, I pre-process the time-series under analysis as follows.

First, the polarisation scores associated with tweets are aggregated at the daily level. Such an aggregation allows me to maintain the granularity of the Twitter data to study short-term variations while minimising potential noise in the data. Aggregating on a monthly or weekly basis could excessively smooth out important fluctuations, potentially neglecting significant trends or shifts in public discourse. Conversely, hourly aggregation might introduce unnecessary noise, capturing too many minor or irrelevant variations. Thus, daily aggregation provides the optimal balance, preserving enough detail to detect meaningful trends while filtering out the noise that could mask broader patterns. Using the scores associated with each tweet, I create a time-series, with each date

[17] I retrieved the most representative words in each tweet's text, representative documents for comparison, number of important words for topic assignment, and a boolean variable to identify whether the tweet is a representative document for its cluster (or, topic).

[18] As an embedding model, I set the "all-MiniLM-L6-v2" model, which is a smaller and faster version of BERT, and it is suitable for large-scale text data and provides efficient embeddings. Available at: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (Accessed on 4 November 2024).

[19] As a representation model, which represents the way text documents are represented before clustering, I set "KeyBERTInspired". This is a representation model that leverages keyword extraction techniques to create document representations, meaning that it can capture the essential semantic information from a given text.

as a polarisation score for each user type under analysis (political actors, news media and public users).

Second, the time-series developed above are made stationary by taking their first difference. Subsequently, I conducted two stationarity tests to confirm stationarity, such as the Augmented Dickey-Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS). The Augmented Dickey-Fuller (ADF) test checks if a time-series is stationary by examining changes in its values over time, assuming non-stationarity as the null hypothesis. If a calculated value is significantly less than zero, one can conclude that the series is likely stationary. The KPSS test is the opposite of the ADF test. It assumes the time-series is stationary and looks for evidence against this assumption. If the calculated test statistic is significantly greater than the critical value, it suggests that the null hypothesis of stationarity is rejected, and the series is likely non-stationary. By employing these tests, I can gain a more comprehensive understanding of a time-series' stationarity. This approach helps mitigate the risk of misinterpretation, as a series might pass one test but fail the other. In my analysis, all time-series became stationary after taking their first difference, enabling me to proceed with further analysis.

### 3.4.2 Employing Vector-Autoregressive Models

To assess my theoretical framework, I employ vector autoregressive (VAR) models on stationary time-series. These models allow me to analyse how the levels of polarisation contained in tweets from various actors (political actors, news media, and general public) influence one another. Specifically, following a large body of previous research (e.g., Barberá et al., 2019; Gilardi et al., 2021; Widmann, 2022; Wlezien and Soroka, 2023; Wlezien, 2024), I fit several VAR models, one for each coded dataset (Republican, Democrat, and Combined coders, as well as GPT-4), via Impulse-Response Functions (IRFs).

VAR models study vectors of variables as autoregressive processes (Box-Steffensmeier et al., 2014; Lütkepohl, 2005; Rachev et al., 2012). In such models, each variable is correlated linearly with its own past lagged values, and those of the other variables in the vector model. Future values of the process are a weighted sum of past and present values, as well as some noise and exogenous variables. Therefore, these models allow researchers to simultaneously study different vector variables (e.g., in this case, tweets' time-series) across time, accounting for potential endogeneity. Based on

theoretical grounds, VAR models, vis-à-vis other time-series analyses (e.g., structural equation approaches), provide several advantages. Researchers can obtain granular statistical results, such as simultaneous analyses between time-series, specific lags, and controlling for endogeneity. Structural equation approaches only study individual coefficients, and deduct the model's dynamic properties (Box-Steffensmeier et al., 2014). Theoretically, the VAR models study the empirical implications of several theories through various variables, identifying the causal chain order of simultaneous effects. In contrast, structural equation approaches rely on single theories for deciding the variables to include in the models (ibidem).

The results of the estimated VAR models are examined using cumulative IRFs. These cumulative functions enable me to study the effect of a certain explanatory variable (e.g., sentiment in tweets, political network, discussion topic) on the outcome variable (e.g., sentiment in replies) after a shock in a given moment, along the time span (Barberà et al., 2019; Lütkepohl, 2005). As such, VAR models enable me, via IRFs, to examine the autoregressive relationships among the time-series under analysis (Box-Steffensmeier et al., 2014; Rachev et al., 2012).

To summarise, based on this discussion, I proceed as follows for each set of polarisation-coded tweets. After fitting the VAR models to the stationary time-series, I determine the optimal number of lags to achieve a balanced bias-variance trade-off, using the Akaike information criterion (AIC)[20] model order selection function (Lütkepohl, 2005). Considering too many lags may lead to high variance (or, overfitting), while too few may lead to high bias (or, under-fitting) (Brandt and Williams 2007: 25, Box-Steffensmeier et al. 2014). Based on previous work on social media, I expect communication to be instantaneous, or fast-paced, with users interacting with each other within short intervals. Moreover, such a window allows me to consider a full week of cyclicality with recurring political events (e.g., press releases, interviews, etc.). Therefore, I set a maximum of 10 lags for the AIC model to find the optimal number of lags. During my analysis, in all cases, AIC selected an optimal lag length of 10 days. Subsequently, following previous research (e.g., Formisano et al., 2024), I analyse the fitted VAR models using cumulative non-orthogonal IRFs, given their lower number of assumptions vis-à-vis orthogonal IRFs. Specifically, to study the relationship of the time-series under analysis, I employ the following equation:

[20]The AIC balances the number of estimated parameters in the VAR model against how well the model fits the given time-series data (Lütkepohl, 2005).

$$\gamma_{i,t} = a_i + \sum_{i,\,t}^{10} \beta_{i,t}\,\gamma_{i,t} + \epsilon_{i,t}$$

In this equation, I set one time-series ($\gamma$) to represent the dependent variable under analysis (predicted polarisation scores) for user type ($i$) in a given day ($t$), with $\varepsilon$ accounting for the error of the model, and with $a_i$ representing a dataset-fixed constant.

To validate such an analytical strategy, I employ several robustness checks. First, beyond cumulative IRFs, I also replicate the analysis using non-cumulative and orthogonal IRFs. Second, individual pairs of time-series are investigated via Granger Causality tests, and only those that pass the 95% confidence threshold are considered to be statistically significant. Below, I discuss the IRFs that pass all robustness checks, while Appendix A12 displays full detailed results.

## 3.5 Results

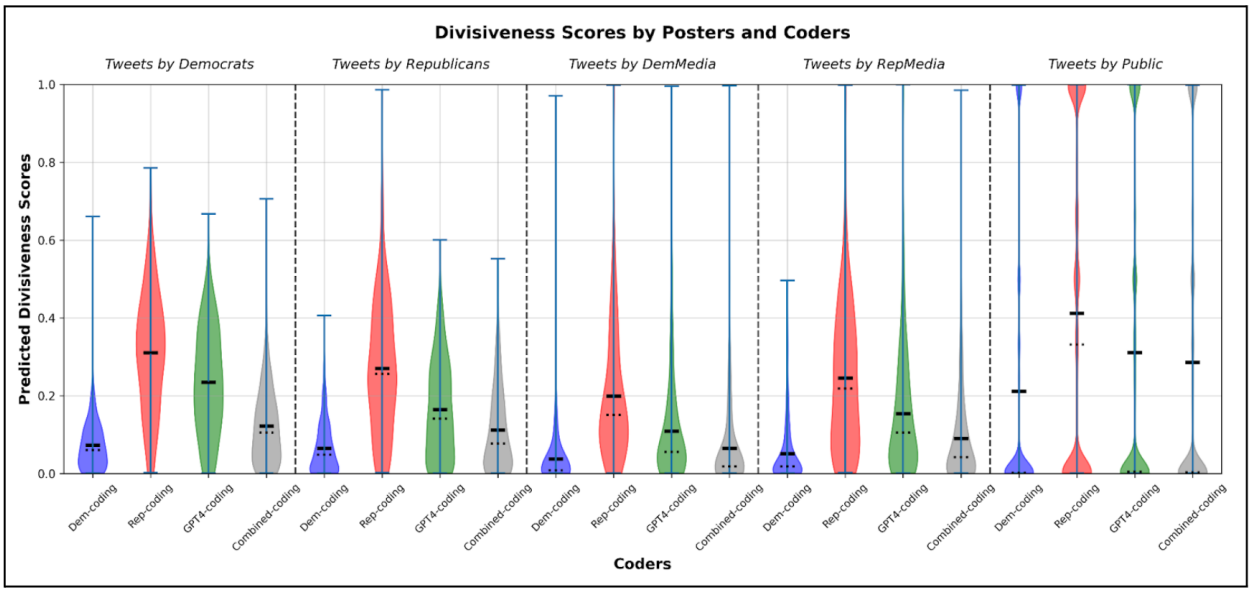
### 3.5.1 Observational Results

In the following subsections, I now provide descriptive statistics regarding the fully coded datasets, accounting for all coders and posters. First, I discuss the distribution of the levels of polarisation, using violin plots. Second, I analyse the time-series showing polarisation over time, studying specific events in US politics.

*Violin Plots*

Figure 1 displays a violin plot that illustrates the predicted divisiveness scores for tweets posted by five user types: Democrats, Republicans, Democrat-leaning Media, Republican-leaning Media, and the Public. Along the x-axis, different coders are represented, including Democrat, Republican, and Combined coders, as well as GPT-based coding. The y-axis shows the predicted divisiveness scores by each classifier, ranging from 0 (lowest) to 1 (highest), indicating how polarising the content of the tweets is perceived to be by each group of coders.

Figure 1: Divisiveness Scores by Tweets' Posters and Coders




*Notes: Divisiveness scores are aggregated (averaged) at the user level.*

Figure 1 reveals that there are no statistically significant differences, as the confidence intervals overlap across different coders and groups of posters. However, there are notable descriptive patterns worth highlighting. For tweets posted by Democrats, the divisiveness scores show some variation across coders. Combined coders assign the highest levels of polarisation, followed by Democrat coders, who perceive a wider spectrum of divisiveness in tweets from their own party, and Republican coders. GPT-4o exhibit the lowest scores, suggesting a more neutral or less sensitive assessment of polarisation in these tweets. Tweets posted by Republicans show similar trends, though with some differences. Combined coders tend to give these tweets the highest divisiveness scores, yet lower divisiveness when compared to Democrat posted tweets. Instead, Democrat coders predict greater variation in the levels of polarisation when compared to Democrat posted tweets, suggesting greater divisiveness when viewed from opposing perspectives. GPT-4o perceives lower divisiveness scores compared to Democrat posted tweets with greater spread of the distribution. In the case of Democrat-leaning (DemMedia) and Republican-leaning (RepMedia) Media, the pattern of predicted divisiveness scores does not always align with expectations. Republicans perceive the highest levels of polarisation in both Media groups, yet the results show RepMedia contain more divisiveness than DemMedia, contrary to expectations. Instead, tweets from DemMedia receive lower than divisiveness ratings from Democrat coders, with RepMedia tweets perceived as more

polarising, which is in line with prior expectations. Similarly, yet marginally, Combined coders and GPT-4o perceive higher polarisation in RepMedia than in DemMedia. Finally, for tweets posted by the Public, the range of divisiveness scores is more evenly distributed across coders. There is still noticeable variation, though it is not as pronounced as with tweets from partisan or media sources. This could suggest that tweets from the general public are perceived as less consistently divisive across political groups. Overall, there is a common pattern with Combined coders finding the highest levels of polarisation, followed by Democrat and Republican coders, with GPT-4o as the least sensitive to polarisation.

*Exploring time-series with Key Events in US Politics*

Figures 2 to 5 display the predicted divisiveness scores associated with each tweet in the dataset over time, capturing shifts in polarisation across different user types (Democrat and Republicans politicians, the public, and partisan media) as coded by four distinct annotative perspectives: Democrat, Republican, Combined and GPT-4o. Each plot includes major political and social events, such as elections, Supreme Court rulings, mass shootings, and significant legislative actions—that likely influenced public discourse. Appendix A9 contains a list of such events. The time-series are plotted using a rolling average of 30 days to smooth short-term fluctuations and highlight broader trends in divisiveness, ensuring that temporary spikes do not hide long-term patterns in the levels of polarisation. Appendix A11 contains detailed descriptions for each plot. Below, I discuss the main trends, key differences, and interpretations that emerge when comparing divisiveness in political tweets as coded by human coders and AI.

In the Democrat-coded plot, divisiveness is initially higher for Republican politicians, especially from 2015 through early 2017, reflecting a perception among Democrat coders that Republican discourse was particularly polarising during the lead-up to and aftermath of Donald Trump's election; this suggests Democrats viewed Republican rhetoric as a driver of polarisation. After early 2017, the gap narrows, with both parties exhibiting similar fluctuations, indicating a convergence in divisive language. However, I record higher spikes for the Democrats when their party was under scrutiny, such as the Russia investigation and Trump's impeachment. Public tweets remain relatively less divisive until early 2019, when a notable rise begins. This shift is likely tied to Twitter's introduction of a new searching function, "Topics", which aimed at helping users find the most relevant tweets to their interests, reinforcing specific interests and creating a feedback loop of content that users were more likely to react to[21] [22]. Such an algorithmic change may have amplified divisive content among the general public. Both Republican and Democrat-leaning media tweets, as coded by Democrats, generally show relatively similar trends yet lower than politicians, implying that media outlets were less likely to use polarising language, though brief surges during major events indicate temporary shifts in tone.

[21] Topics on X. Available at: https://help.x.com/en/using-x/follow-and-unfollowtopics#::text=On%20mobile%2C%20tap%20Topics%20in%20your%20profile, on%20the%20button%20next%20to%20the%20topic (Accessed on 31 October 2024).

[22] The "Topics" feature allowed users to follow curated themes rather than an individual account.

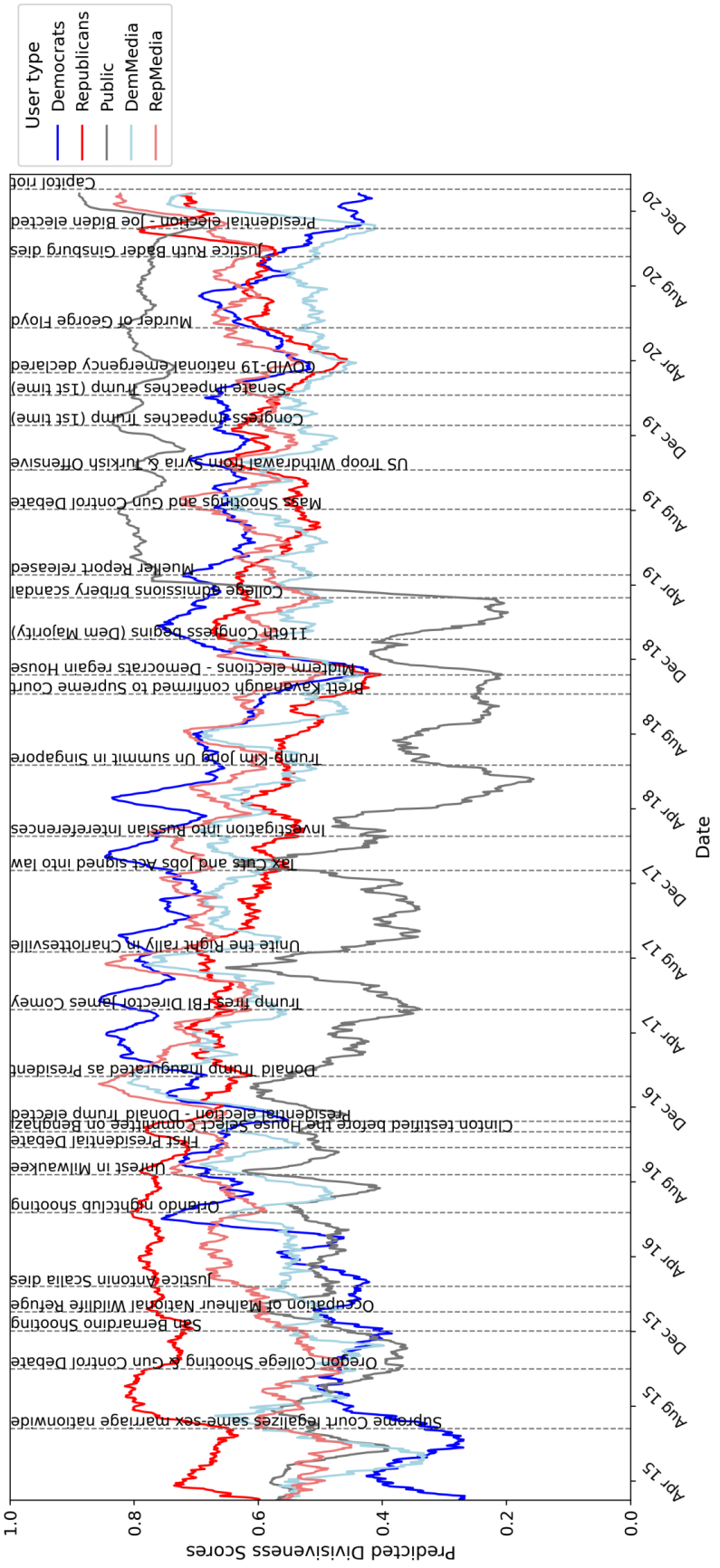



Figure 2: Divisiveness Democrat-coded Tweets Over Time

The Republican-coded plot provides similar results to the Democrat-coded tweets. Overall, Democrats are seen as more divisive than Republicans, except before November 2016 and after Biden's election in 2020. Republican coders perceive Democratic rhetoric as more inflammatory, particularly during events threatening Republican interests, such as the months following Trump's inauguration and Comey's firing. Republican peaks are associated with conservative flashpoints, indicating that Republican coders are sensitive to events that challenge their values. As per Democrat-coded tweets, the public's divisiveness remains low until early 2019, after which it rises sharply, possibly due to Twitter's "Topics" feature being introduced. Media tweets here are also generally lower than politicians in divisiveness but with marginally higher baseline scores and sharper spikes than in the Democrat-coded plot, indicating that Republican coders perceive media coverage as more divisive, especially during crises.

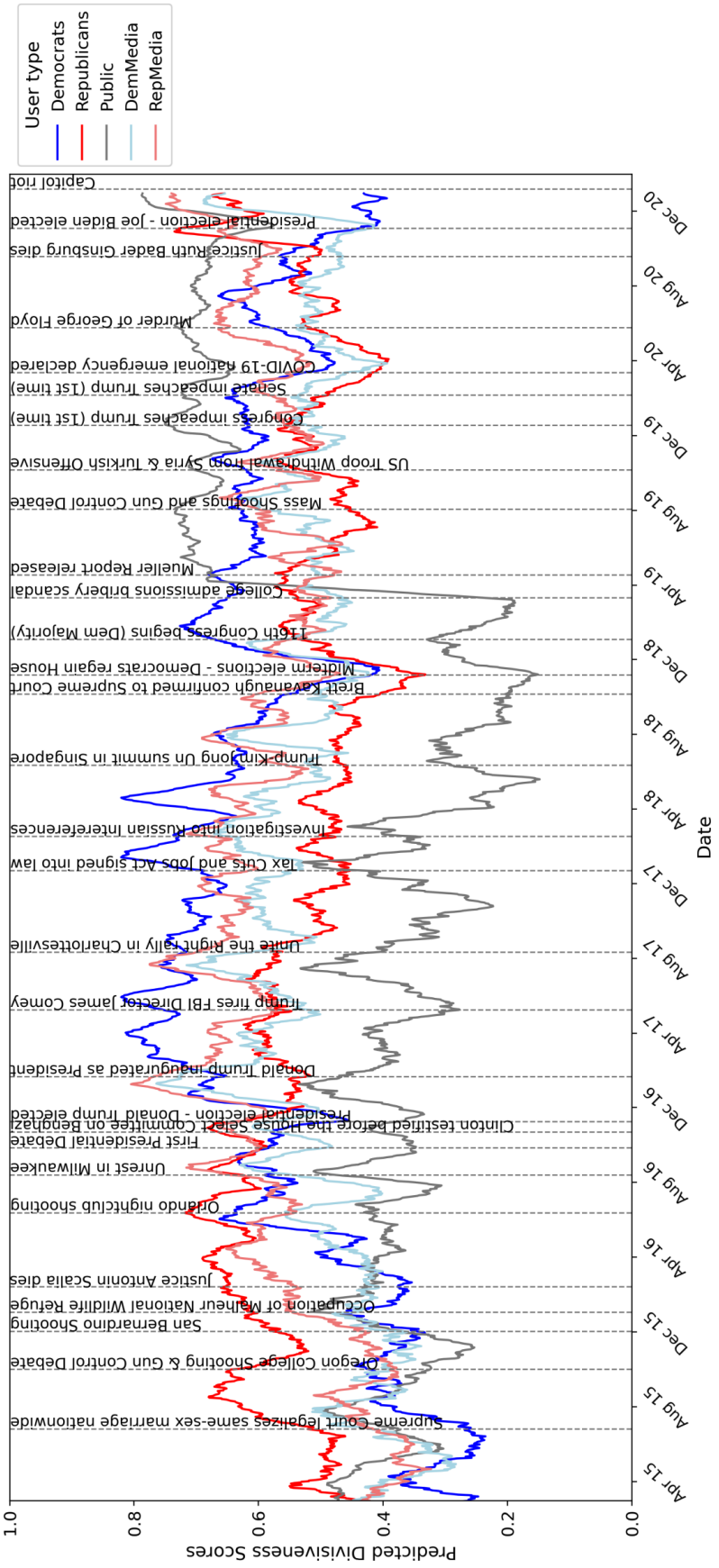



Figure 3: Divisiveness Republican-coded Tweets Over Time

The Combined-coded plot integrates both partisan perspectives and closely resembles the Democrat-coded trends, with Republicans seen as more divisive leading up to the 2016 election, after which the gap closes and both parties' divisiveness becomes more synchronised. However, Democrat politicians display higher levels of divisiveness overall. Democratic and Republican spikes relatively align with high-profile national events, indicating that both parties' rhetoric becomes more polarising during crises. The public shows a strong spike during the Milwaukee unrest and a sustained rise after April 2019, implying that divisiveness became more widespread beyond elite circles. Media divisiveness fluctuates in response to major events, such as Trump's inauguration, suggesting media outlets generally maintain a less divisive tone except during national crises.

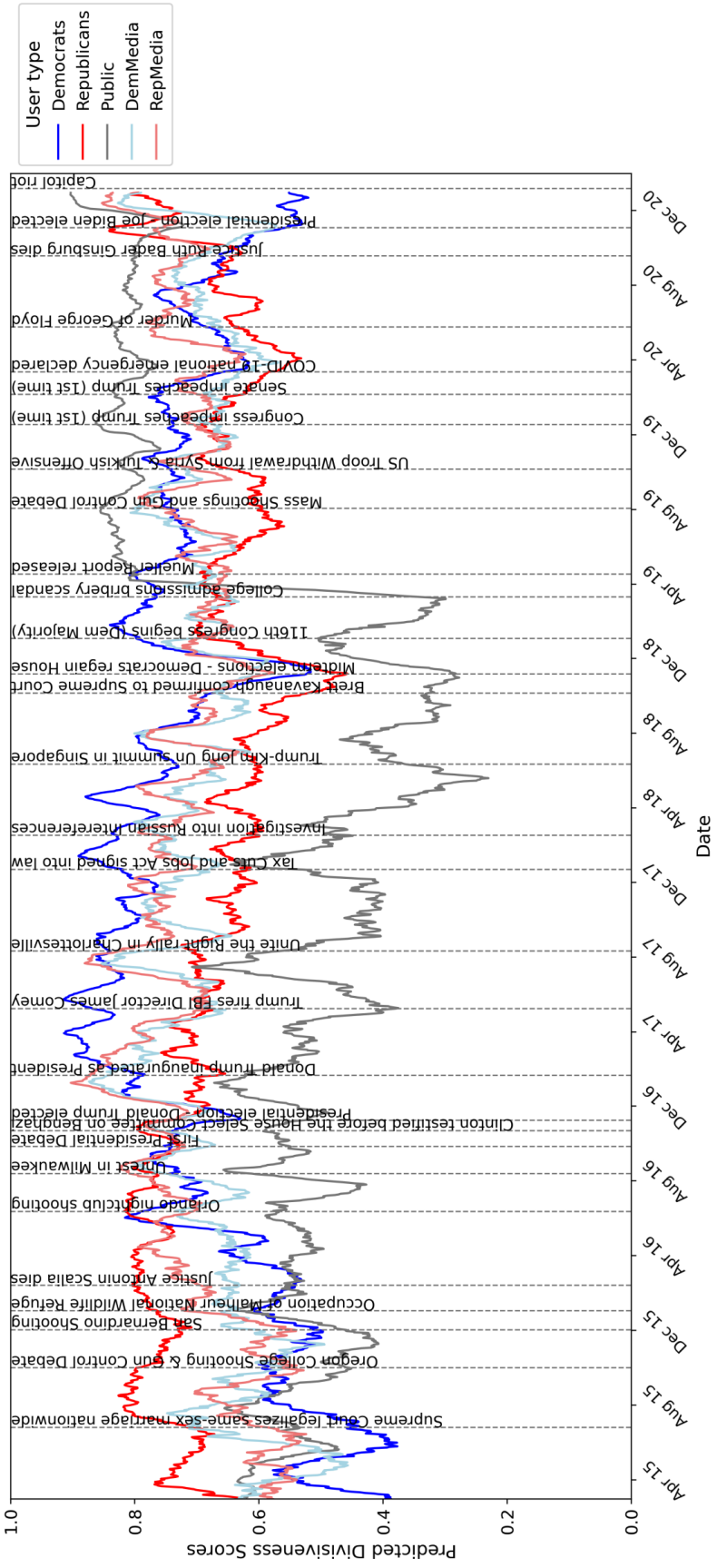



Figure 4: Divisiveness Combined-coded Tweets Over Time

The AI-coded (GPT-4o) plot reveals a different dynamic, with much lower divisiveness scores overall. The trends seen above remain, with Republican politicians being more divisive from early 2015 to late 2016, Democrat politicians surpassing them significantly after that, a trend that persists until about April 2019. After this point, public divisiveness rises to match Republican partisan levels, with a resurgence during and after the 2020 election, indicating that polarisation is no longer confined to political elites. GPT-4o identifies a general upward trend in divisiveness across all groups, suggesting a broadening of polarising discourse on Twitter. Media-aligned users show much lower levels of polarisation significant, even in response to major events.

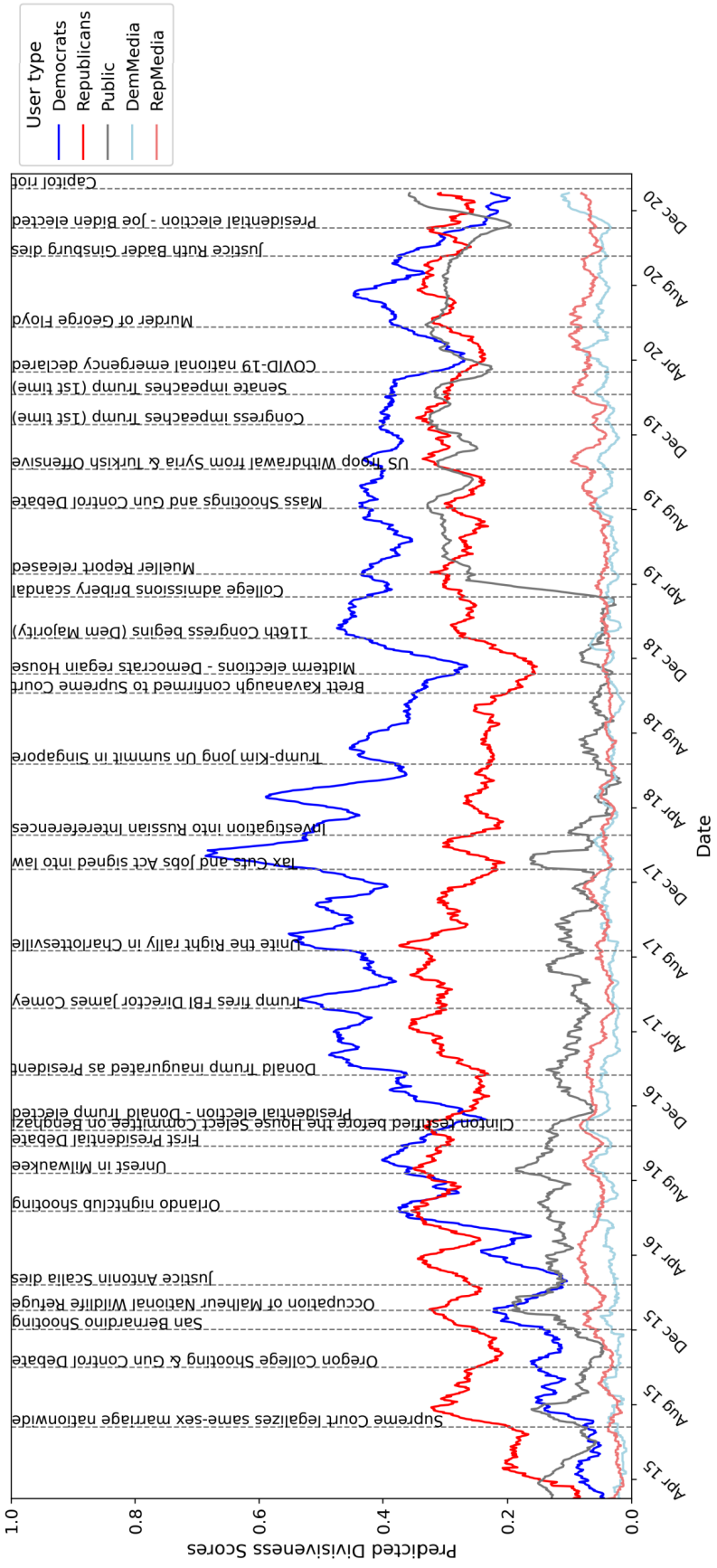



Figure 5: Divisiveness GPT-coded Tweets Over Time

Across all coding schemes, divisiveness among the public is relatively low and stable until early 2019, when it rises sharply, likely due to algorithmic amplification on Twitter, suggesting a democratisation of divisiveness. Media accounts are generally less divisive than politicians but display notable spikes during major crises, implying that media polarisation is event-driven. All coders identify the same key events as flashpoints for divisiveness, but the perceived source and intensity vary, underscoring how coder alignment and context shape interpretations of polarisation.

### 3.5.2 Discussing Correlational Results

Having provided descriptive analyses of the coded tweets regarding their overall distribution and over time, I now move to correlational results to establish associations regarding the levels of polarisation among the users under analysis. Table 2 summarises the IRFs output and Granger Causality Tests, displaying the coders, relationship of variables, Granger p-value, and time lags. The table only presents the statistically significant relationships that have passed the Granger test ($p<0.05$), the remaining non-significant results are presented in Appendix A12.

Table 2: Summary of the IRFs output and Granger Causality Tests

| Coders | Relationship | Granger P-Value | Time lags | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Democrat** | **Dem Div —>Rep Div** | 0.0083 | - | - | 0.065 | - | - | - | - | - | - | 0.054 | 0.08 |
| | **Rep Div —>Dem Div** | 0.04782 | - | 0.068 | - | - | - | - | - | - | - | - | - |
| | **Rep Div —>Public Div** | 0.00064 | - | 0.081 | - | - | - | - | - | 0.061 | - | - | - |
| | **Rep Div —>RepMedia Div** | 0.00959 | - | 0.068 | - | - | - | - | - | - | - | - | -0.045 |
| **Republican** | **Dem Div —>Rep Div** | 0.0367 | - | - | 0.056 | - | - | - | - | - | - | 0.054 | 0.057 |
| | **Dem Div —>DemMedia Div** | 0.01293 | - | 0.063 | - | - | - | - | - | - | - | 0.073 | 0.056 |
| | **Rep Div —>Public Div** | 4.00E-05 | - | 0.053 | - | - | -0.053 | - | - | - | - | -0.076 | -0.056 |
| | **Rep Div —>RepMedia Div** | 0.02594 | - | 0.061 | - | - | - | - | - | - | - | -0.051 | - |
| | **Public Div —>Rep Div** | 0.00603 | - | 0.073 | - | - | - | - | - | - | 0.046 | - | - |
| | **DemMedia Div —>RepMedia Div** | 0.00348 | - | 0.068 | 0.073 | 0.056 | - | - | - | 0.044 | - | - | - |
| **Combined** | **Dem Div —>Rep Div** | 0.00803 | - | - | 0.071 | - | - | - | - | 0.054 | - | 0.07 | 0.072 |
| | **Dem Div —>Public Div** | 0.03148 | - | 0.082 | 0.093 | 0.075 | - | - | - | - | - | 0.079 | 0.07 |
| | **Dem Div —>DemMedia Div** | 0.02108 | - | 0.101 | - | - | - | - | - | 0.064 | - | 0.069 | 0.056 |
| | **Dem Div —>RepMedia Div** | 0.02922 | - | 0.067 | - | - | - | - | - | - | 0.097 | - | - |
| | **Rep Div —>Dem Div** | 0.00962 | - | 0.06 | 0.046 | - | - | - | - | - | - | - | - |
| | **Rep Div —>Public Div** | 0.01406 | - | 0.085 | - | - | - | - | - | - | - | - | - |
| | **Public Div —>Rep Div** | 0.01562 | - | - | - | - | - | - | - | - | 0.04 | - | 0.044 |
| | **Public Div —>DemMedia Div** | 0.00503 | - | 0.064 | - | - | - | - | - | 0.073 | - | - | - |
| | **RepMedia Div —>Dem Div** | 0.01164 | - | - | - | - | - | - | 0.054 | - | - | -0.04 | -0.035 |

Notes: All relationships concern the levels of polarisation (“Dev”) for each poster: Democrat politicians (“Dem”), Republican politicians (“Rep”), general Public (“Public”), Democrat-leaning Media users (“DemMedia”) and Republican-leaning Media users (“RepMedia”). Each lag corresponds to a day.

Next, I examine these relationships across the different coder groups, only considering the statistically significant estimates ($p < 0.05$).

For tweets coded by Democrat annotators, there is a mild positive relationship between Democrat and Republican politicians' levels of polarisation (Dem Div → Rep Div) at lag 2 (0.065), lag 9 (0.054), and lag 10 (0.08). This suggests that Democrat politicians' divisive rhetoric may influence subsequent Republican tweets, being associated with increased polarisation after a delay. This delayed response indicates that Republicans may take time to craft responses that match the level of polarisation, or that the discourse gradually intensifies over time. In contrast, there is a positive relationship when switching the correlation arrow from Republican to Democrat politicians (Rep Div → Dem Div) at lag 1. This suggests that Democrats are instead fast at responding to polarising cues from the Republicans, yet such a relationship is short-lived, as it does not repeat over time. Additionally, Republican politicians appear to influence public divisiveness (Rep Div → Public Div) at lag 1 (0.081) and lag 7 (0.061), showing that their rhetoric has an immediate effect, which persists until the end of the week, on public polarisation. Similarly, Republican divisiveness positively correlates with Republican-leaning media (Rep Div → RepMedia Div) at lag 1 (0.068), reinforcing the idea that political rhetoric shapes media discourse. However, this effect becomes negative at lag 10 (-0.045). One possible explanation is that divisive rhetoric may initially drive media polarisation, but as time passes, factors such as audience fatigue, shifting political priorities, or strategic recalibration by media outlets may contribute to a reduction in divisive coverage. This pattern suggests a dynamic relationship where media responses to political rhetoric evolve over time rather than maintaining a consistent trajectory.

For tweets coded by Republican annotators, several significant relationships emerge. There is again a positive relationship between Democrat and Republican politicians' levels of polarisation (Dem Div → Rep Div) at lag 2 (0.056), lag 9 (0.054), and lag 10 (0.057). This corroborates the previous results regarding Democrat politicians' divisive rhetoric may influence subsequent Republican tweets, leading to increased polarisation after a delay. There is a positive relationship between Democratic polarisation and Democratic-leaning media divisiveness (Dem Div → DemMedia Div) at lag 2 (0.063), lag 9 (0.073), and lag 10 (0.056), implying that Democrat politicians' rhetoric influences media narratives in a relatively short time. Republican politicians, affect public polarisation (Rep Div → Public Div) at lag 1 (0.053), but this effect turns negative at lag 4 (-0.053), lag 9 (-0.076),

and lag 10 (-0.056). This pattern suggests that while Republican divisiveness initially influences public discourse, it may also contribute to a counterreaction over time, where public sentiment moderates instead of escalating. Furthermore, Republican politicians influence Republican-leaning media polarisation (Rep Div → RepMedia Div) at lag 1 (0.061) reinforcing the role of partisan media in amplifying divisive rhetoric. However, as above, the effect turns negative at lag 9 (-0.051). Interestingly, public divisiveness also influences Republican politicians (Public Div → Rep Div) at lag 1 (0.073) and lag 8 (0.046), indicating a feedback loop where public discourse can shape political rhetoric. Lastly, Democratic-leaning media influences Republican-leaning media (DemMedia Div → RepMedia Div) at lags 1 (0.068), 2 (0.073), 3 (0.056), and 7 (0.044) suggesting an interplay between opposing media narratives.

The combined coder results reveal a positive relationship between Democrat and Republican politicians' divisiveness (Dem Div → Rep Div) at lag 2 (0.071), with a sustained effect at lag 7 (0.054), lag 8 (0.07), and lag 9 (0.072). This reinforces previous findings from Democrat and Republican coders. Moreover, it may furnish evidence about reciprocal polarisation, where divisive rhetoric from one side prompts a response from the other, reinforcing partisan divides over time. Additionally, Democrat politicians' rhetoric influences public divisiveness (Dem Div → Public Div) at lags 1 (0.082), 2 (0.093), and 3 (0.075), and sustained at lags 9 (0.079) and 10 (0.07), suggesting that Democratic polarisation resonates quickly with the public and has a lasting effect. Democrat politicians also influence Democrat-leaning media (Dem Div → DemMedia Div) at lags 2 (0.101), 7 (0.064), 9 (0.069) and 10 (0.056), showing that media actors react to narratives pushed by aligned politicians over time. Similarly, Democrat politicians influence Republican-leaning media (Dem Div → RepMedia Div) at lags 1 (0.067) and 8 (0.097). The effect at lag 1 suggests an immediate counter-narrative or reactive framing, while the stronger effect at lag 8 implies a longer-term adaptation, possibly as narratives gain traction or become politically advantageous to engage with. This pattern highlights the interplay between partisan media ecosystems, where opposing media outlets do not merely operate in isolation, but actively respond to and shape discourse driven by political figures from both sides. In contrast, Republican-leaning media appears to influence Democrat politicians (RepMedia Div → Dem Div) at lag 7 (0.054), but this effect turns negative at lag 9 (-0.04) and lag 10 (-0.035), suggesting an initial polarising response followed by later moderation. Republican politicians in turn influence Democrats' levels of polarisation (Rep Div → Dem Div)

at lags 1 (0.06) and 2 (0.046), however, this effect is short-lived. Such a relationship reinforces previous findings of the reciprocal nature of partisan polarisation, where divisive rhetoric from one party provokes a reaction from the opposing party, but this response is not sustained over time. Additionally, there is a reciprocal relationship between Republicans' rhetoric and the public (Rep Div$\rightarrow$ Public Div; Public Div $\rightarrow$ Rep Div). Republicans positively influence the public at lag 1 (0.085), but this effect is short-lived. In contrast, the levels of polarisation in the public influences Republicans later at lags 8 (0.04) and 10 (0.044). This delayed effect suggests that while Republican rhetoric quickly shapes public polarisation, the public's divisiveness takes longer to feed back into Republican political discourse. This could indicate that politicians initially drive polarisation, but over time, they respond to shifts in public sentiment, potentially adjusting their rhetoric to align with the evolving attitudes of their base. The public also has an effect on Democrat-leaning media (Public Div $\rightarrow$ DemMedia Div) at lags 1 (0.064) and 7 (0.073). This indicates that Democrat-leaning media quickly reacts to shifts in public polarisation, possibly by amplifying prevailing narratives or adjusting coverage to align with audience sentiment. The effect at lag 7 implies a longer-term adaptation, where media outlets may gradually integrate public divisiveness into their framing and reporting strategies over time.

Interestingly, the time-series data representing the levels of polarisation in tweets, as coded by GPT-4o, reveal no significant relationship between the users under analysis. This could be explained by the fact that GPT-4o identified generally low levels of polarisation across the tweets, as seen in Figure 1. As a result, the lack of a clear relationship between users suggests that the content analysed was relatively neutral or non-controversial, leading to minimal polarisation.

### 3.5.3 Assessing Hypotheses

As previously argued, my theoretical framework on the spread of polarisation is three-fold. First, politicians are likely to follow the public rather than leading. Second, news media users will have a catalyst role in such interactions. Third, cross-cutting interactions between users with different partisanships can occur, increasing polarisation among them overall, or decreasing on common/shared political grounds. Below, I analyse the findings in light of each hypothesis.

The results provide mixed support for Hypotheses 1 and 2, especially regarding the dynamics involving the media and public influences on politicians. Several observations suggest that while

the public and media impact political discourse, the response from politicians can vary based on partisan leanings.

Public influence on politicians: considering the Republican coders' coded tweets, I find further support for the notion that the public influences politicians. There is a positive relationship between the public's and Republican politicians' levels of polarisation at lags 1 and 8. Considering the combined coders' coded tweets, I find further support for this notion. The relationship between the public's polarising language and Republican politicians is positively significant at lags 8 and 10, indicating that the public may indeed precede and influence the language Republicans use later on. This aligns with my hypothesis, suggesting that, over time, Republican politicians follow shifts in public discourse, particularly in polarising rhetoric. Considering Democrat politicians, there is no evidence showing that Democrat politicians are directly led by public discourse. Instead, Democrat politicians seem to respond more strongly to cues from Democrat-leaning media, which I analyse below.

Media influence as a catalyst: There is strong evidence that media plays a critical role in shaping political discourse, often acting as a catalyst for polarisation. Democrat politicians influence Democrat-leaning media (Dem Div $\rightarrow$ DemMedia Div) at several lags, highlighting a relationship where political figures and media actors mutually reinforce each other's divisive discourse. This suggests that Democrat-leaning media serve as both a reflector and amplifier of Democratic political rhetoric. Furthermore, Democrat-leaning media appear to influence Republican-leaning media (DemMedia Div $\rightarrow$ RepMedia Div) at lags 1, 2, 3, and 7 (Republican coders) and lags 1 and 8 (Combined coders), suggesting an interplay between opposing media ecosystems. This could indicate that Republican-leaning outlets react quickly to Democrat-leaning media's narratives, possibly to counteract them or to frame their own partisan responses. Such cross-cutting interactions highlight how partisan media do not operate in silos but engage with each other's rhetoric, further contributing to polarisation. Republican-leaning media appears to influence Democrat politicians (RepMedia Div $\rightarrow$ Dem Div) at lag 6, though this effect reverses at later lags (9, 10), when considering combined coding. This suggests that, while Republican-leaning media may initially provoke a polarising response from Democrats, this influence wanes over time, potentially due to strategic recalibration or audience fatigue. Similarly, Republican-leaning media are influenced by Republican politicians (Rep Div $\rightarrow$ RepMedia Div) at early lags, reinforcing the idea that partisan media act as an amplifier

of political rhetoric. However, this effect turns negative at later lags, suggesting that while initial coverage might be highly divisive, factors such as public response, changing political priorities, or media fatigue may contribute to a reduction in polarising content over time. The public influences Democrat-leaning media (Public Div $\rightarrow$ DemMedia Div) at lags 1 and 7 (combined coding), indicating that media outlets react to shifts in public sentiment. Thus, public discourse shapes media narratives, which in turn may further amplify polarisation. The longer-term effect at lag 7 implies that media responses to public divisiveness evolve over time, rather than being purely immediate reactions. Nevertheless, beyond combined coding, such a relationship is not supported.

Politicians' influence on the public: Although this relationship speaks against my hypotheses, it is important to note that there is evidence of politicians influencing the public in using polarising language, as observed in all subsets of coded tweets. For the Democrat-coded dataset, there is a negative relationship between Republican politicians and the public's levels of divisiveness at lags 1 and 7. For the Republican-coded dataset, there is a positive relationship only at lag 1, turning negative at lags 4, 9 and 10. Considering combined-coded tweets, Republicans influence the public at lag 1 only. In contrast, Democrats may precede and influence the language the public uses on a longer time-span, as seen at lags 1, 2, 3 as well as 9 and 10.

Taken together, I find partial support for Hypotheses 1 and 2. The analyses partially support the hypothesis that the media acts as a catalyst, but politicians often lead and follow the public, rather than simply following. Politicians' responses are not uniform and appear shaped by partisan alignment, with Republicans more influenced by public discourse and Democrats by media.

The findings support Hypothesis 3 of cross-polarisation interactions, highlighting distinct patterns based on which side initiates polarising language. Given their statistical significance, I analyse the VAR-results related to Democrat, Republican, and Combined coded tweets.

Cross-Partisan Influence on Polarisation: for Democrat-coded tweets, a mild positive relationship emerges between Democrat and Republican polarisation at certain lags (lags 2, 9, and 10), suggesting that Democrats' divisive rhetoric prompts delayed Republican responses, fuelling polarisation. Conversely, Republicans positively influence Democrats to use more divisive language at lag 1 only. Considering Republican-coded tweets, the results confirm a positive relationship at lags 2, 9, and 10, implying that when Republicans escalate polarisation, Democrats may moderate their rhetoric,

possibly to counterbalance the discourse. However, the combined coder results reveal a sustained positive relationship between Democrat and Republican polarisation across multiple lags (2, 7, 9 and 10), reinforcing the idea that polarisation tends to escalate through reciprocal influence over time.

Partisanship and Media Influence: partisan media tends to reinforce polarisation within their ideological sphere, while also influencing opposing politicians. Considering both Democrat and Republican coders, Republican politicians positively influence Republican-leaning media in the short (lag 1) but negatively in the long run lag 10 (Democrat coders) and 9 (Republican coders). According to Republican and Combined coders, Democrat politicians positively affect Democrat-leaning media at both short and long-term lags (1, 7, 9, and 10, for both coders). Additionally, Democrat-leaning media positively influence Republican-leaning media, intensifying their polarising rhetoric, and providing evidence for cross-polarisation interactions over the media sphere at lags 1, 2, 3 and 7 (Republican coders) and at lags 1 and 8 (combined). Similarly, combined coder results show a positive relationship between Republican-leaning media and Democrat politicians' polarisation at lag 6, suggesting that both parties react to opposing partisan media narratives. However, such a relationship turns negative at lags 9 and 10.

These findings highlight the reciprocal nature of partisan polarisation, with both Democrats and Republicans influencing each other's rhetoric over time. Partisan media further amplifies this dynamic, reinforcing polarisation within ideological spheres while also shaping opposing narratives.

## 3.6 Discussion

The goal of this paper has been to investigate the temporal and directional dynamics of polarisation among politicians, news media, and the public in US politics (2015-2020) on Twitter.

The descriptive analyses show that there are no statistically significant differences in the distribution of polarisation among groups, however some patterns have emerged. Combined coders consistently assign the highest divisiveness scores, followed by Democrat and Republican coders, while GPT-4o produces the lowest scores, indicating a more neutral stance. Democrat coders perceive a wider range of divisiveness in their own party's tweets, but still lower rates than those assigned to either Republican politicians or media. Republican coders rate Democrat tweets less divisive than

as perceived by Democrat coders, and they perceive Republican-leaning Media as more divisive than Democrat-leaning Media, contrary to expectations. GPT-4o and Combined coders also find Republican-leaning Media more polarising. Public tweets show more balanced divisiveness ratings across coders, suggesting less consistent polarisation.

The time-series observational analyses show that Republican politicians used the highest levels of polarisation from early 2015 to late 2016, which are followed by the Democrat politicians who displayed the highest levels from late 2016 to early 2019, being surpassed by the public afterwards. The plots paint an interesting picture, revealing that there are some key events in the spread of polarisation among all posters. For example, it shows noteworthy spikes related to the occupation of Malheur National Wildlife Refuge (January 2016), Trump's elections (November 2016), mass shootings and unrest (San Bernardino 2015, Orlando 2016, Milwaukee 2016), and Trump's impeachment proceedings, Russian interference in the 2016 elections (February 2018) and college admission scandals (April 2019), US troops withdrawal from Syria (October 2019), Senate impeachment of Trump (February 2020), murder of George Floyd (May 2020), and Capitol Hill riot (January 2021). Importantly, Twitter's platform updates in April 2019 may have had significant impacts on the public's levels of polarisation. Finally, tweets posted by the media displayed the lowest levels of polarisation among all posters.

The VAR analyses explored the relationship between polarising language used by Democrats, Republicans, the public, and news media users. Overall, the results show competing findings when accounting for who posts and reads the social media content, suggesting a complex interplay between these groups, with varying degrees of influence and response. The analyses partially support the hypothesis that the media acts as a catalyst in line with previous work by Wlezien (2024) and Wlezien and Soroka (2024). Specifically, media coverage tends to mirror existing public preferences and policy decisions more than it independently drives them. However, partially consistent with previous literature (e.g., Barberá et al., 2019; Hawkins et al., 2020; Tromble, 2018), politicians both lead and respond to public opinion, shaping discourse on key issues while also adapting their positions in response to public sentiment, rather than simply following the electorate. Additionally, the results confirm cross-polarisation interactions, showing that both parties shape each other's rhetoric over time. This finding contributes to previous works by Hong and Kim (2016) and Bright (2018). In particular, the analyses show that Democrats often drive Republican polarisation, while Republicans

occasionally elicit moderation from Democrats, creating a fluctuating cycle of political division. These findings underscore the reciprocal and evolving nature of online polarisation, emphasising the need to account for temporal dynamics in future research.

Some limitations of this study warrant consideration. The current analytical pipeline is designed for studying polarisation in public discourse in US politics, and thus it is based on a majoritarian party system, however this may limit its portability to multi-party systems. In countries with multiple political parties, the interactions among political actors, voters, and media can be more complex, potentially requiring extra training phases for the classifier, in which new data can be added to not only encode texts in different languages, but also encode party and news media usernames to capture the nuances of polarisation. Future research should explore the adaptability of this analytical framework in various electoral contexts, particularly in systems where coalition politics and ideological diversity play significant roles (for example, Germany and Italy). Although the current pipeline uses social media discourse as its primary data source, the flexible architecture of the classifiers and time-series approach allows for exploration of alternative data sources. This adaptability is especially relevant given the variations in time-series results observed across different classifiers (see Figures 2 to 5), which can be further investigated. For instance, researchers could apply classifiers to full news media articles retrieved from digital newspaper databases such as Factiva[23], Lexis-Nexis[24], and Retriever[25], enabling a longitudinal analysis of how media narratives contribute to polarisation over time. Additionally, politicians' press releases and parliamentary speeches could provide deeper insights into the strategic framing of issues and responses to public sentiment. Finally, another potential avenue could be employing these classifiers to study open-ended survey questions, where the public is asked about opinions about opposing or their own favourite parties, over time. In this sense, the different partisanship of the coders can be significant to truly capture mechanisms and how they are perceived.

Having developed four algorithms to measure polarisation (Paper 1), and having examined these dynamics over time (Paper 2), this thesis now turns to the propagation of polarisation through social networks. Paper 3 investigates the network structures, constructing networks from mentioned users and analysing their connections and node characteristics to map how polarised discourse spreads.

[23] Available at: https://global.factiva.com/ (Accessed on 6 November 2024).
[24] Available at: https://www.lexisnexis.com/en-us (Accessed on 6 November 2024).
[25] Available at: https://www.retrievergroup.com/ (Accessed on 6 November 2024).

# 4 The Network Dynamics of Online Polarisation in the US

**Abstract**

The role of online political communication among politicians is widely acknowledged as a key component of democratic deliberation. While previous research has explored the rise and patterns of polarised rhetoric on social media, less is known about how these communication behaviours shape the structural dynamics of political networks. This paper investigates the network structures of US politicians engaging in political discourse, identifying key network features linked to polarised language. Specifically, I employ hierarchical exponential random graph models (HERGMs) to predict ties between users based on contextual factors, such as their party affiliation. I also consider the degree of polarised language present in their tweets, as determined by supervised machine learning classifiers based on human coders and artificial intelligence. Descriptive analyses reveal that Democrat politicians hold stronger connections to other powerful nodes in the network than Republicans. HERGM analysis shows high party-based homophily, with most connections occurring within political groups and these ties being mutual and generally low in polarising language. In contrast, cross-party mentions are rare and much more likely to involve divisive or conflictual language, highlighting that interactions between political groups are typically divisive rather than cooperative. These insights highlight the structural and linguistic drivers of online political polarisation.

**Keywords:** Affective Polarisation, Machine-Learning, Networks, US Politics, Twitter.

## 4.1 Introduction

Social media, particularly Twitter, has fundamentally reshaped political communication by allowing political actors to disseminate information and engage directly with the public, blurring boundaries between formal and informal discourse (Benkler, 2007; Yoon and Park, 2014). Twitter is now a key source of political information and a widely used communication channel (Smith, 2011), revealing the structures of political alliances and oppositions and making it essential for analysing polarisation and coalition-building. Studying political elites on Twitter is crucial for understanding these dynamics, as the platform offers a direct, unfiltered space where politicians can share views, influence discourse, and bypass media gatekeepers (Tucker et al., 2017; Jungherr, 2016; Enli and Skogerbø, 2013; Theocharis et al., 2016). At the same time, according to many, Twitter

contributes to disinformation and ideological polarisation, with elites using it to reinforce partisan narratives and microtarget public opinion (Tucker et al., 2018). The literature on Twitter's impact on democratic discourse presents a theoretical and methodological puzzle. Some argue that Twitter fosters ideological insularity by amplifying like-minded voices and filtering out dissent, reinforcing polarisation (Hong and Sung, 2016; Sunstein, 2017). In contrast, other studies highlight the platform's potential to expose users to diverse viewpoints and bridge partisan divides, suggesting a more pluralistic effect (Barberá et al., 2015; Colleoni et al., 2014; Vaccari et al., 2016). However, further empirical evidence complicates both perspectives: while Twitter's architecture allows for both partisan reinforcement and ideological crossover, political elites and users overwhelmingly engage within their own ideological circles, with retweeting and following behaviours reinforcing social clustering (Vergeer, 2020; Boutet et al., 2012; Himelboim et al., 2013; Conover et al., 2011; Esteve Del Valle and Borge Bravo, 2018a; Gruzd and Roy, 2014).

The nature of engagement further complicates the picture. While mentions can bridge ideological camps, they rarely indicate cooperation or deliberation; instead, politicians often mention opponents to provoke or signal opposition, not to foster dialogue (Del Vicario et al., 2016; Qiu et al., 2019; Hegelich and Shahrezaye, 2015). Research shows that it is ideological divergence—not just incivility—that fuels polarising interactions online, with cross-party mentions typically serving as rhetorical conflict rather than engagement (Avalle et al., 2024). Experimental evidence reinforces this: exposure to opposing views on social media tends to harden, not soften, ideological positions, especially among Republicans, aligning with theories of motivated reasoning and affective polarisation (Bail et al., 2018; Nyhan and Reifler, 2010; Taber and Lodge, 2006; Bail, 2015). While classic intergroup contact theory suggests dialogue can reduce prejudice (Pettigrew and Tropp, 2006), the anonymity and performative nature of Twitter often undermine these effects, encouraging incivility and entrenchment instead (Bail et al., 2018). Although only a small portion of social media content is anonymous, most abusive or uncivil content comes from these anonymous accounts or posts, while incivility still occurs—though less often—even when real identities are used. Thus, anonymity can fuel online aggression, but it is not the only cause of toxic discourse across platforms (Rösner and Krämer, 2016).

These previous works raise key theoretical and methodological questions about how polarisation flows and under what conditions cross-cutting interactions can occur or reduce division. Building

on these debates, this paper addresses the following research question:

> *How do structural network dynamics and the content of political communication jointly shape polarisation among US politicians on Twitter?*

To investigate these dynamics, in this study, I employ hierarchical exponential random graph models (HERGMs) to examine the Twitter discourse among US politicians. HERGMs are social network analysis tools, which can model complex network structures by considering both individual-level attributes (e.g., party affiliation, number of mentions, levels of polarisation) and structural, or higher-order dependencies (e.g., propensity for a politician to mention another politician, as well as their likelihood of reciprocal interactions) that influence engagement patterns. Beyond network structures, the content of political communication plays a pivotal role in shaping public perception and policy discourse. To this end, my study integrates textual analysis methods to measure polarisation within the content of tweets, while controlling for the posters' party affiliation. The levels of polarisation are measured via supervised machine learning, trained on human coders (i.e., Republican, Democrat, Combined) and AI (i.e., GPT-4o), as explained in papers 1 and 2. By assessing ideological leanings and the levels of divisive rhetoric employed in these communications, I aim to capture the extent and nature of polarisation in the digital narratives crafted by politicians. This dual focus on both the network-structure and content of interactions provides a well-rounded analysis of political discourse in the digital age.

The findings for this study are three-fold. First, centrality measure investigations reveal Democrat politicians are more connected to influential people in the network, while in other ways their connections are similar to Republicans. Second, HERGM results indicate that within political groups, people tend to connect with those who share their views, and these connections are often mutual and contain low levels of polarising language. In essence, it is more likely to see Democrats mentioning other Democrats than Republicans, and when this happens it typically reflects supportive or unifying interactions rather than divisive rhetoric, reinforcing internal group cohesion. While users tend to connect with others who share their political views, interactions between different political groups often involve more polarising language, reflecting both the divide and attempts to connect across these divides. In other words, mentions across party lines—such as a Democrat mentioning a Republican—are much rarer and more likely to involve polarising or conflictual

language, highlighting the divide between political groups and the contentious nature of cross-party engagement. Finally, by integrating network analysis with textual measures of divisiveness and comparing coder effects, this study shows that structural and discursive mechanisms jointly reinforce partisan clustering and conflict-driven cross-party interactions. Notably, while human coders show strong consensus in identifying within-group patterns, they diverge from AI models in capturing the intensity of polarisation in cross-group ties, underscoring how methodological choices shape our understanding of elite digital networks. These findings highlight the structural constraints, linguistic patterns, and methodological factors shaping online political polarisation.

## 4.2 The Network Dynamics of Online Polarisation

In this paper, I examine the relevance of affective polarisation to online behaviour and network dynamics. Foundational studies on affective polarisation argue that individuals in political discussions express positive sentiments toward in-group members and negative ones toward out-groups (Iyengar et al., 2012; Iyengar and Westwood, 2015; Mason, 2015, 2018). In practice, this translates into more positive language and engagement directed at co-partisans, and negative language used against ideological opponents. These affective dynamics are especially visible on social media, where positive interactions typically occur within like-minded groups, and cross-group interactions are more often hostile or critical (Avalle et al., 2024; Del Vicario et al., 2016; Fraxanet et al., 2024; Qiu et al., 2019).

To investigate dynamics, researchers have turned to social network analysis (SNA), which has been employed to understand how political connections are formed (Esteve Del Valle and Borge Bravo, 2018; Esteve Del Valle et al., 2022; Song, 2015), to predict political outcomes (Wojcik, 2019), and to analyse the nature of political relationships (Gonzalez-Bailón and Wang, 2016; Gruzd and Tsyganova, 2015). Fraxanet et al. (2024) note that SNA has been especially useful in studying factionalism and elite polarisation in digital spaces, with research focusing on contexts such as the UK House of Commons during Brexit (Intal et al., 2021), coalition building in parliamentary democracies (Fontan and Altafini, 2021), and polarisation in the US Congress and House of Representatives (Aref and Neal, 2020; 2021; Capozzi et al., 2023). Recent work on newer platforms such as Bluesky offers further insights. Quelle and Bovet (2025) studied the interaction network of approximately five million users from February 2023 to May 2024. Bluesky, as a decentralised plat-

form, allows users to customise algorithms and self-host their data. Despite its structural differences from Twitter, the platform still exhibited high clustering, short path lengths, and ideological polarisation—particularly in discussions regarding contentious issues like the Israel–Palestine conflict. While misinformation was minimal and centre-left news sources dominated, significant issue-based polarisation persisted, underlining the need to study polarisation even in decentralised environments. However, these studies often overlook the content and tone of the language used in these political interactions.

The presence of both political peers and the public on Twitter creates two key embedded contexts for politicians: network boundaries and public visibility (Yoon and Park, 2014). As defined by Granovetter (1985), embeddedness refers to how interpersonal ties shape behaviour, with actors embedded in networks facing distinct constraints and opportunities (Moody and White, 2003; Yoon and Park, 2014). An example of these ties is socio-communicational embeddedness, which can be seen in how an individual's actions are shaped by their friend's network (Monge and Contractor, 2003). Following social network analysis, political actors on Twitter can be seen as embedded within dynamic, interdependent structures, wherein mutual and dyadic ties shape access to information, influence, and visibility—central processes in networked environments (Borgatti et al., 2009). As Borgatti and colleagues emphasise, the structure of ties between actors, rather than their individual attributes, plays a significant role in shaping outcomes such as diffusion, coordination, and social capital. Moving back to Twitter, interactions are public for any other user to see and easily traceable, these relational mechanisms are amplified, reinforcing patterns of homophily, clustering, and centrality. Politicians' behaviours are thus not only influenced by direct connections but also by their position within the larger network and by structural tendencies like reciprocity and transitivity, which foster dense clusters and influence dynamics of power and visibility within the platform.

Given its characteristics and considering the constraints of pre-internet-based network analyses, Twitter has been a central focus in this area. Previous works studying ties among politicians have studied bill-co-sponsorship networks, where politicians' choices reflect both policy content and their relationships with fellow supporters, creating selective and persuasive ties (Fowler, 2006; Kessler and Krehbiel, 1996; Wilson and Young, 1997). Legislators use co-sponsorship to signal support, driven by electoral incentives or strategic positioning (Mayhew, 1974; Campbell, 1982; Kessler and Krehbiel, 1996). While studies have examined group patterns and ideological clustering (Panning,

1982; Pellegrini and Grant, 1999; Talbert and Potoski, 2002), less is known about which legislators receive the most support, despite co-sponsorship being seen as a marker of leadership (Hall, 1992; Krehbiel, 1995; Schiller, 1995). Taken together, the boundary seeks to signal other politicians the content of the bill and express some persuasive power over individual politicians (Fowler, 2006; Wilson and Young, 1997). In contrast, Twitter imposes no such selective boundaries—its shared platform turns all participating politicians into potential network connections, not only those making co-sponsored bills (Yoon and Park, 2014). This fosters a denser, more transitive network structure, consistent with social network tendencies noted by Holland and Leinhardt (1972) and Wasserman and Faust (1994). As previously discussed, unlike traditional legislative settings, Twitter enables politicians to interact in a highly visible, real-time public arena, bypassing media gatekeepers and blurring the lines between formal and informal communication (Benkler, 2007; Smith, 2011; Yoon and Park, 2014). The platform's architecture facilitates not only broad dissemination of information but also strategic use of interactive features—such as mentions and retweets—to signal alliances, reinforce partisan narratives, and engage in public debate (Arnaboldi et al., 2022; Cherepnalkoski et al., 2016).

However, scholars are divided on its effects: some argue it amplifies ideological segregation and echo chambers (Hobolt et al., 2024; Hong and Sung, 2016; Sunstein, 2017), while others find that it increases exposure to disagreement (Barberá et al., 2015; Colleoni et al., 2014; Vaccari et al., 2016). Empirical studies show that political communication tends to stay within ideological boundaries. Vergeer (2020) found that during the 2012 Dutch elections, candidates mainly interacted with their own party or ideologically adjacent parties. Boutet et al. (2012) observed similar patterns in the 2010 UK general election, where party members primarily retweeted their own. Himelboim et al. (2013) found that users discussing the 2010 US midterms engaged mostly with like-minded others, forming homogeneous clusters.

Further examining Twitter's political discourse, Flamino et al. (2023) studied nearly a billion tweets from the 2016 and 2020 US presidential elections. Their analysis combined political bias data with fact-checking sources to measure the circulation of biased content. They observed a decline in fake and extremely biased news over time, but a growing ideological divide and stronger echo chambers, as users became less likely to engage with opposing viewpoints.

Despite overall trends toward homophily, Twitter networks—particularly those involving politicians—show some diversity in viewpoints. For instance, Hong and Sung (2016) found that US House members selectively engaged with political content aligned with their views, but not exclusively. Crucially, the structure of interaction matters. Several studies highlight that the type of Twitter activity—whether following, retweeting, or mentioning—shapes the level of ideological diversity in networks. Following and retweeting tend to reinforce homophily (Conover et al., 2011; Esteve Del Valle and Borge Bravo, 2018a; Esteve Del Valle et al., 2020; Yoon and Park, 2014), whereas mentions are more likely to bridge ideological divides. Conover et al. (2011) and Ogawa et al. (2013) found that while users mostly retweet like-minded individuals, they mention opponents more often. Most online networks are sparse and contain structural holes that prevent cross-interactions; therefore, the presence of bridging ties and active information brokers is essential for diffusion to occur (Gonzalez-Bailón and Wang, 2016). Investigating these gaps, Yoon and Park (2014) have applied exponential random graph model and regression methods to Twitter data, showing that Korean politicians' follower networks were ideologically clustered, yet mention networks enabled more cross-party interaction. This suggests that party affiliation may play a predictive role in the formation of mention ties. Building on this evidence, I formulate the following hypothesis.

**Hypothesis 1:** Party affiliation predicts mentions (edges) between users.

Few studies have explored how different types of Twitter connections relate to cross-ideological communication (Avalle et al., 2024; Bail et al., 2018; Hegelich and Shahrezaye, 2015). While Evans et al. (2014) show that politicians use Twitter for a range of purposes, their analysis does not link these behaviours to specific communication channels, and their reliance on hand-coding limits the scalability of their findings. Similarly, Ausserhofer and Maireder (2013) investigated the Twitter networks of Austrian politicians but did not distinguish between types of mentions. These studies primarily focus on the existence of mentions, without considering the content of those interactions—which could provide insight into how politicians express support or disagreement, and how messaging patterns differ across ideological lines. This gap raises an important question: might divisive or polarising content actually drive cross-party mentions, rather than reflect genuine deliberative engagement?

In this regard, scholarship by Avalle et al. (2024) analysed 500 million comments from eight

platforms—Facebook, Gab, Reddit, Telegram, Twitter, Usenet, Voat, and YouTube—over three decades. They found that toxicity was consistent across platforms and time, but that the presence of polarised viewpoints had a stronger impact on the evolution of online discussions than toxicity itself. Hostility often emerged in response to contrasting political perspectives. These results are in line with previous work on how users with opposing views engage more negatively with each other (Del Vicario et al., 2016; Qiu et al., 2019). Further evidence of this pattern appears in Hegelich and Shahrezaye (2015), who study the tweets posted by German MPs over two months before the 2013 General Elections, from September 2013 to December 2013. To analyse the data, they use a combination of networks (e.g., Kolmogorov–Smirnov test) and sentiment analysis, which include the number of followers and friends (e.g., users that are followed and follow back), as well as looking at the tweets' text and mentioned users. Their findings reveal that politicians are more likely to use negative language when addressing members of opposing parties directly, underscoring the tendency for cross-party communication to take a more hostile tone. Similarly, previous literature shows how cross-cutting interactions accentuate perceived differences between groups and increase their commitment to preexisting beliefs (Bail, 2015; Bail et al., 2018; Nyhan and Reifler, 2010; Taber and Lodge, 2006). Notably, Bail et al. (2018) investigated whether exposure to opposing political views on Twitter can reduce polarisation. In a field experiment, they incentivised Democrats and Republicans to follow a bot for one month that shared content from the opposing political ideology. The results showed that Republicans became more conservative after following a liberal bot, while Democrats exhibited a small, statistically insignificant shift toward more liberal views after following a conservative bot. These findings suggest that, rather than bridging ideological divides, exposure to cross-cutting political content on social media may sometimes intensify existing beliefs and deepen polarisation.

Another line of research (e.g., Pettigrew and Tropp, 2006) suggests that contact between opposing groups can help counteract stereotypes that often form when positive interactions are lacking, as noted by Bail et al. (2018). This previous work also shows that intergroup contact can promote deliberation and political compromise (Huckfeldt et al., 2004; Mutz, 2002). However, these studies have primarily focused on direct, in-person interactions between individuals from rival groups. This paper instead focuses on online conversations between politicians from the opposing political camp on a social media platform. Hence, it is unclear whether such online interactions foster the same

kind of mutual understanding as face-to-face encounters. Nevertheless, as argued by Bail et al. (2018), there is evidence to claim that the anonymity and lack of non-verbal cues on social media actually encourage incivility, potentially intensifying conflict rather than reducing it. This previous work allows me to develop the following hypothesis.

**Hypothesis 2:** Higher polarisation is positively associated with cross-party interactions.

Building on these hypotheses, my contribution aims to demonstrate that conversations within political groups tend to exhibit strong party-based homophily. While cross-party interactions on Twitter do occur, they are often characterised by divisive or antagonistic language. Rather than fostering openness or deliberative engagement, such exchanges are more likely to reinforce polarisation by emphasising sharp distinctions and confrontational discourse.

## 4.3 Data

This paper analyses the dataset developed in the previous two chapters, which includes tweets posted by politicians, manually annotated by their political affiliations, as well as the levels of polarisation used in their texts. This section explains the construction of directed networks, beginning with the definition of network boundaries and the characteristics of nodes and ties to be analysed.

To plot and visualise the communication network among the politicians, I employ the following procedure. First, I construct a directed graph to represent user interactions, where each node corresponds to a Twitter user, and directed edges signify mentions of other users. To account for political affiliations, each node is assigned a colour attribute—blue for Democrats and red for Republicans. The analysis focuses exclusively on tweets posted by politicians, as these users provide the complete set of required information (e.g., party affiliation).

In line with previous work by Castanho Silva and Proksch (2022), the analysis only includes politicians who have mentioned the official handle of other politicians. By focusing on official handle mentions, the study ensures that the resulting network accurately reflects intentional, traceable communication between politicians, minimising ambiguity and enhancing the reliability of the interaction data.

The network does not include public and news media users. The former, public users, are simply

excluded, given current Twitter's restrictions that do not allow researchers to estimate their ideological leaning. Excluding tweets from news media accounts, despite their party-leaning tags, is essential in this network analysis to maintain analytical clarity and accurately interpret the dynamics of political interactions. News media often serve as information disseminators rather than active participants in political discourse, and their inclusion could introduce biases that obscure the patterns of engagement among politicians. Additionally, media accounts may exhibit different networking behaviours compared to individual politicians, such as broader outreach and content distribution strategies, which could distort the analysis of interpersonal political relationships.

Moreover, I exclude politicians who do not mention other politicians, ensuring that the network consists only of actively engaged users. This exclusion is necessary because isolated nodes, or those without outgoing mentions, do not contribute to the network structure in a meaningful way and could distort degree-based analyses.

In these networks, the levels of polarisation associated with each tweet have been predicted by Democrats, Republicans, and Combined coders, as well as GPT-4o, using the procedure described in detail in papers 1 and 2. In building the networks, the averaged predicted levels of polarisation in the text are used as a characteristic of the ties. This procedure allows each connection to reflect not only the existence of interaction (or mention) between users, but also the degree of polarisation associated with that interaction. Therefore, it is now possible to analyse user interactions and relationships within the defined network boundaries while ensuring that all included nodes participate in the political discourse, making the network a more accurate representation of political interactions on Twitter.

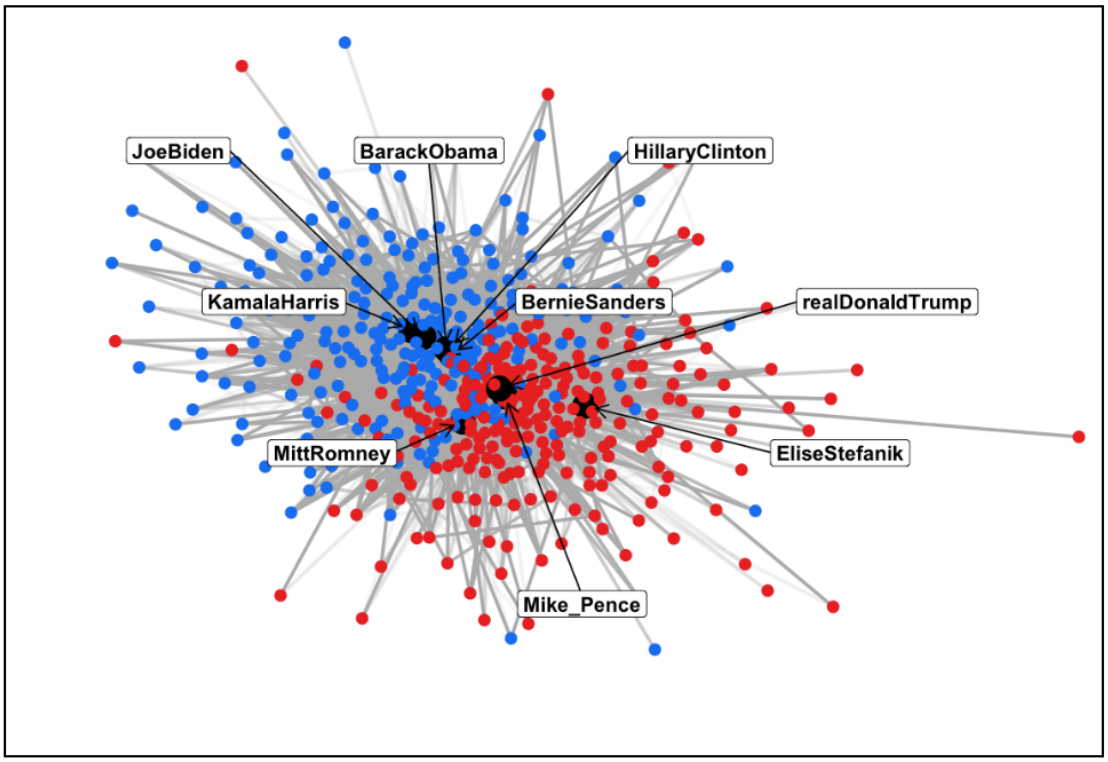
Table 1 summarises the key network statistics, following the above-mentioned network-building procedure.

Table 1: Summary of Key Network Statistics

| **Statistics** | **N** |
|---|---|
| Nodes | 508 |
| Edges | 5,783 |
| Democrats | 253 |
| Republicans | 255 |

Finally, the network is visualised using the Fruchterman-Reingold (FR) layout (Fruchterman and Reingold, 1991), a force-directed algorithm that positions nodes based on attractive and repulsive forces. This layout helps reveal the structural properties of the network by clustering highly connected nodes while maintaining some separation between different groups. In this plot, the nodes represent individual politicians and directed edges indicating mentions between them. Nodes are colour-coded by party affiliation, with red for Democrats and blue for Republicans. The structure reveals having a dense core within each cluster and suggests frequent intra-party interactions, while the limited connections between the two clusters highlight the polarised nature of political communication. Specifically, the few nodes positioned between and across the clusters represent individuals engaging in cross-party interactions, which are relatively rare compared to within-party mentions.

Figure 1: Network visualisation




*Notes: This network visualisation shows interactions among politicians, where each node represents a user, coloured by party affiliation (red for Republicans, blue for Democrats). Connections indicate mentions between users. The Fruchterman-Reingold layout emphasises strong within-party clustering, and limited cross-party interactions.*

## 4.4 Methods

This section is divided into two complementary parts. First, I present some descriptive statistics of the different networks, such as examining network density, centrality measures (e.g., degree, betweenness, closeness), and community detection. Second, I explain the operation of hierarchical exponential random graph models (HERGMs) and detail the robustness checks that will be performed to ensure the reliability of the results. This evaluation step includes assessing model convergence, testing for multicollinearity among the two edge covariates, and conducting goodness-of-fit tests.

### 4.4.1 Descriptive analyses

For the descriptive analyses, I consider the following centrality measures based on Borgatti and Everett (2006) and Borgatti et al. (2018): In-degree measures the number of incoming connections a node receives, reflecting its popularity or influence within the network. Out-degree represents the number of outgoing connections a node initiates, indicating its level of activity in engaging with others. Betweenness centrality quantifies how often a node serves as a bridge along the shortest paths between other nodes, highlighting its role in facilitating information flow. Closeness centrality measures how efficiently a node can reach all others in the network, with higher values indicating shorter average path lengths. Eigenvector centrality extends the concept of degree centrality by considering not just direct connections but also the influence of those connections, meaning nodes connected to highly influential others receive higher scores. The clustering coefficient reflects the degree to which a node's neighbours are interconnected, capturing the extent to which a given node is embedded within a tightly-knit community.

### 4.4.2 Correlation analyses

To conduct correlation analyses, I employ HERGMs, which are an extension of simpler ERGMs, to analyse network structure, focusing on within-cluster and between-cluster connectivity.

Exponential random graph models (ERGMs) are statistical frameworks used to analyse and model the complex structures of networks by considering the various factors that influence the formation of edges between nodes (An, 2016; Lusher et al., 2013). ERGMs allow researchers to assess

how specific configurations, such as the presence of triangles (indicative of triadic closure) or the overall density of the network, contribute to the observed network structure. They are typically estimated using Monte Carlo Maximum Likelihood Estimation (MCMLE), which relies on Monte Carlo Markov Chains (MCMCs) to sample networks for estimation.

However, these processes pose both computational and conceptual challenges (An, 2016). With large networks, ERGMs can be slow, particularly for large networks, as MCMCs take time to reach the target distribution (ibidem). They require significant computer memory and software capable of handling large data, which exceeds the capacity of standard commercial computers. Furthermore, as network size increases, the assumption that actors have full knowledge of the network when deciding connections becomes increasingly unrealistic, complicating the estimation process (ibidem). Finally, but crucially for this paper, traditional ERGMs may face limitations when dealing with networks that exhibit hierarchical or multi-level structures, where dependencies exist not only at the dyadic level but also across different levels of the hierarchy. Specifically, in a simple ERGM, edges are randomly formed between nodes. Each pair of nodes has an independent, fixed probability of being connected. Hence, this type of modelling assumes that every pair of nodes has an equal chance of being connected, regardless of their properties, leading to a network that tends to be more homogeneous in terms of connectivity.

To overcome these limitations, Clauset et al. (2006) introduced the Hierarchical exponential random graph models (HERGMs), which extend traditional ERGMs by incorporating hierarchical structures, making them particularly suitable for analysing large and sparse networks. In network analysis, hierarchy refers to the tendency of networks to have a pattern where smaller groups of nodes (vertices) cluster together, and these groups then form larger groups, which can continue on in a similar way (Clauset et al., 2006). This creates a structure that looks like a "fractal," where the same pattern repeats at different levels, from the smallest parts of the network to the whole network itself (ibidem: 3). These structures are different from hierarchical clustering, which groups network nodes by combining them in steps to form a hierarchy. In such clustering, it remains unclear whether these structures reflect the actual network or if they are simply a result of the way the algorithms work. This blending of network's properties with the algorithms' effects is a problem that I aim to address. A hierarchical network, as defined by Clauset et al. (2006), is one that naturally breaks into groups, which then break into smaller subgroups, continuing until we reach individual nodes.

This structure is typically shown as a tree or dendrogram, on which I further elaborate below.

HERGMs are capable of capturing both assortative and disassortative structures, as well as any combination of the two, across all scales and parts of a network (Clauset et al., 2008). The first term, assortative mixing, refers to the likelihood of nodes to connect with other nodes that are similar in some way, such as nodes with high degrees (many connections) connecting to other high-degree nodes, and low-degree nodes connecting to other low-degree nodes. In contrast, disassortative mixing describes the opposite pattern, where nodes with different degrees have a greater tendency to connect with each other. In other words, high-degree nodes often link with low-degree nodes, resulting in a network where highly connected nodes are dispersed among less connected ones. The HERGMs construct a probabilistic binary tree of network vertices, which represents the probability of connections and the hierarchical relationships between them. By accounting for multi-level dependencies and nested groupings within data, HERGMs can effectively model complex interactions in expansive networks where direct connections between nodes are relatively infrequent. Moreover, such networks' hierarchical properties can improve the accuracy of predicting missing links, which is crucial in various applications such as social networks, biological networks, and the internet (Clauset et al., 2008). These predictions are the result of incorporating both the above-mentioned global network structure and local connectivity patterns. HERGMs are crucial when studying intricate systems, such as social media interactions among politicians, where engagement patterns are influenced by various hierarchical factors (ibidem). For these reasons, HERGMs are considered to be key approaches for hierarchical inference (Clauset et al., 2008).

In this paper, I fit an HERGM to capture both structural and attribute-based mechanisms influencing tie formation in the network. The model accounts for basic edge formation probability (edges), while incorporating mutuality to capture the tendency for reciprocal connections (mutual). It also includes terms for triadic closure (triangle) and further refined by the geometrically weighted edgewise shared partner (gwesp), which models clustering tendencies in the network. The model also integrates node attributes on party affiliation, assessing whether individuals are more likely to connect with others from the same group. Additionally, edge covariates for divisiveness and weight influence are included to examine whether these factors impact tie formation. The hierarchical structure, with two blocks, allows for clustering effects, capturing within- and between-group differences in network formation. Finally, the model runs on multiple cores, and to ensure reproducibility it uses

a fixed random seed across the different pre-processing and analytical steps, as well as for parallel computations[26].

### 4.4.3 Robustness checks for correlations

As robustness checks, I run multicollinearity tests among covariates and conduct goodness-of-fit (GOF) tests. For the first, I check whether edges' attributes (e.g., weights, divisive scores) correlate.

For the goodness-of-fit assessment, I provide three evaluations to ensure that the model accurately represents both local connection patterns and global network structure: (1) geodesic distance, (2) in-degree and (2) out-degree distributions. First, the geodesic distance measures the shortest path lengths between nodes in the network, offering insight into the overall connectivity structure. A good fit in this aspect suggests that the model accurately captures the extent to which nodes are indirectly connected through intermediaries. If the observed geodesic distribution aligns well with the simulated results, it indicates that the model successfully reproduces the network's reachability and structural cohesion. However, significant discrepancies—such as an over- or underestimation of long paths—could imply that the model fails to account for key structural properties like network fragmentation or small-world effects. Second, the in-degree distribution examines how incoming ties are distributed across nodes, which is crucial in directed networks where certain nodes may act as central hubs. A mismatch between the observed and simulated in-degree distributions could indicate that the model fails to capture hierarchy or preferential attachment effects. The out-degree distribution assesses the pattern of outgoing connections from nodes, providing insight into how activity levels vary across the network. If the simulated and observed distributions differ significantly, it may suggest that the model does not properly account for variation in node-level connectivity behaviour.

To operationalise my GOF assessment on the fitted hierarchical ERGM, I specify several network statistics to compare between the observed and simulated networks. These statistics include degree distribution (degree), measuring the number of ties each node has; geometrically weighted dyad-wise shared partners (gwdsp), accounting for clustering effects by emphasising dyads with different numbers of shared partners; geometrically weighted degree (gwdegree), down-weighting nodes

[26]Specifically, I ensure consistent parallel random number generation by setting a fixed L'Ecuyer-CMRG seed, matching the random seed used for other tasks, to maintain reproducibility.

with extremely high degrees to prevent overfitting; triangle counts (triangles), measuring the number of closed triads in the network, capturing local clustering tendencies and reinforcing transitivity effects; edgewise shared partners (esp), accounting for the number of shared neighbours between directly connected nodes to see how connections contribute to local clustering and the formation of densely connected substructures; and geodesic distance (distance), measuring the shortest path length between node pairs, reflecting the overall structure of the network in terms of how easily information or influence spreads. The number of simulations is set to 2000 to ensure robust comparisons, and a fixed random seed is used for reproducibility.

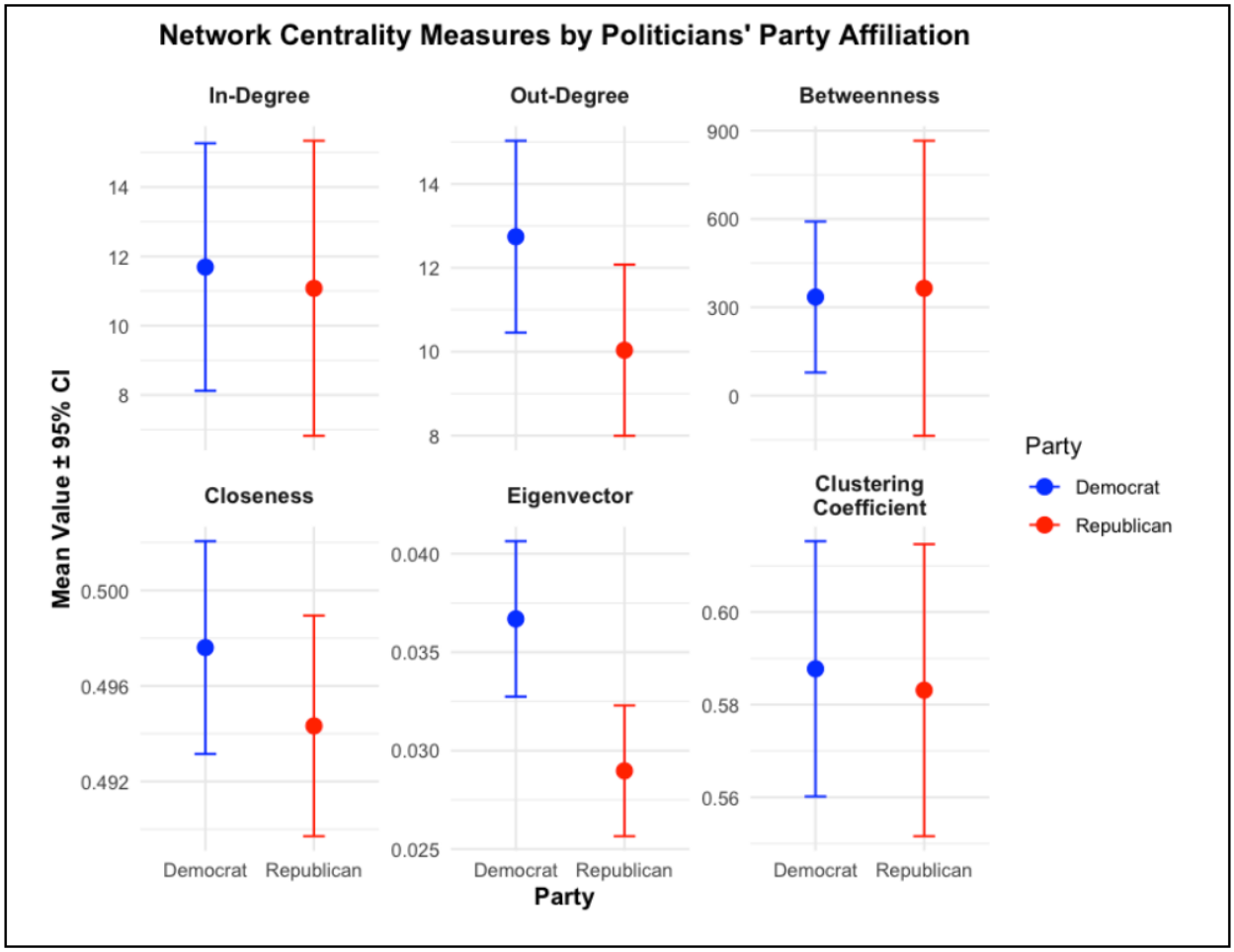
## 4.5 Results

This section presents the results of the network analyses, including both descriptive measures and HERGM outcomes. It details all robustness checks performed, along with interpretations of the findings, to provide a comprehensive understanding of the network structures and relationships identified.

### 4.5.1 Descriptive Findings

Figure 2 presents the mean centrality measures (in-degree, out-degree, betweenness, closeness, eigenvector, and clustering coefficient) for Democrats (blue) and Republicans (red), along with standard errors. Below, I discuss them in turn.

Figure 2: Network Centrality Measure by Politicians' Party Affiliation




*Notes: This figure presents the mean centrality measures (in-degree, out-degree, betweenness, closeness, eigenvector, and clustering coefficient) for Democrats (blue) and Republicans (red), along with standard errors (e.g., the standard deviation of the sample divided by the square root of the sample size).*

Figure 2 shows that Democrat politicians are more consistently connected to other influential nodes in the network, positioning them within more powerful or central communication clusters. In other words, this suggests that Democrats occupy strategically advantageous positions within the political communication network, where they have greater access to and interaction with other highly influential actors. Examining in-degree centrality, Democrats and Republicans show similar values (approximately 11.7 and 11 respectively) with substantially overlapping confidence intervals, suggesting no statistically significant difference in how many connections they receive, contrary to the previous assertion. For out-degree centrality, Democrats display a higher mean (~12.6) compared

to Republicans (~10.1) with marginally overlapping confidence intervals, indicating Democrats may initiate more connections, though this difference is only weakly significant. For betweenness centrality, Republicans show a slightly higher mean but with dramatically wider confidence intervals extending from negative values to approximately 850, compared to Democrats' narrower range (100-600). This extensive overlap indicates no statistically significant difference in bridging roles between parties, contradicting earlier claims about Republicans' strategic positioning. Regarding closeness centrality, Democrats exhibit a marginally higher mean (~0.498) than Republicans (~0.494) with slightly overlapping confidence intervals, suggesting Democrats may be somewhat more efficient at reaching others in the network—the opposite of what was previously claimed. The eigenvector centrality measure shows the clearest partisan distinction, with Democrats (~0.037) significantly higher than Republicans (~0.029) and non-overlapping confidence intervals. This statistically significant finding indicates Democrats are more consistently connected to influential nodes within the network. Finally, clustering coefficients for both parties are remarkably similar (Democrats ~0.59, Republicans ~0.585) with substantially overlapping confidence intervals, contradicting the previous assertion that Republicans form more tightly-knit communities.

### 4.5.2 HERGMs Estimations

I now discuss the HERGMs' results to investigate how connections form within and between different groups. Essentially, these analyses show the extent to which users are likely to interact with others in their own group versus users in other groups. Within-cluster connections refer to how often and how closely users only interact with other users from their own side—like Democrats mostly talking to other Democrats. A strong effect here means most interactions stay within the same group. Between-cluster connections, on the other hand, show the likelihood of users to reach out to those on the other side—like Democrats talking to Republicans. If this estimate is low, it means groups mostly interact among themselves. If it is high, it means there is evidence for cross-group interaction. This analysis assesses whether political conversations on Twitter mostly happen within echo-chambers, or if there are dialogues across party lines. Appendix A13 contains the full detailed results.

Table 2: HERGMs' within-cluster estimation

| Coders | Edges | Mutual | Triangle | Party Affiliation | Divisiveness | Weights |
|---|---|---|---|---|---|---|
| **Democrat** | -5.866*** | 0.438*** | 0.113*** | 0.627*** | -0.215*** | -12.157*** |
| **Republican** | -5.866*** | 0.440*** | 0.113*** | 0.626*** | -0.226*** | -12.177*** |
| **Combined** | -5.865*** | 0.441*** | 0.113*** | 0.633*** | -0.210*** | -12.867*** |
| **GPT-4o** | -5.839*** | 0.485*** | 0.123*** | 0.651*** | -0.180*** | -12.353*** |

*Notes: Models included controls for the following structural terms: gwesp (0.5 for Democrat-, Republican-, and Combined-coded tweets; 0.3 for GPT-4o). P-values < 0.1 (*), 0.05 (**), 0.01 (***).*

Table 2 displays HERGMs' within-cluster estimation. The results show that, within political groups, connections are relatively rare but tend to be mutual and formed between like-minded individuals, with little evidence of divisive language. This aspect indicates that intra-party ties are shaped more by shared views than by polarising discussions. The edges' coefficient is highly negative and statistically significant (-5.866, $p < 0.001$), indicating that within-cluster connections are relatively rare, reinforcing the sparsity of the network. The mutual term is positive and statistically significant (0.438 to 0.485, $p < 0.001$), suggesting a strong tendency for reciprocity, meaning that if one node links to another, the other node is likely to reciprocate the connection. The triangle term is positive and statistically significant (0.113 to 0.123, $p < 0.001$), suggesting a tendency for transitive closure within clusters—that is, if node A is connected to node B and node B is connected to node C, then node A is more likely to also be connected to node C. This indicates a cohesive substructure, where triadic relationships are more prevalent than what would be expected by chance. The party affiliation coefficient is consistently statistically significant (0.627 to 0.651, $p < 0.001$) across all coders, indicating that nodes tend to form connections with others from the same political party, highlighting strong homophily effects. The divisiveness term is negative and statistically significant (-0.215 to -0.180, $p < 0.001$), suggesting that divisiveness may slightly discourage tie formation within clusters, though the effect is less pronounced than other factors. The weights' parameter is consistently large in magnitude and highly negative across all clusters (-12.157 to -12.867, $p < 0.001$), reinforcing the notion of overall network sparsity. These strong negative values suggest that, even within clusters, the formation of ties remains selective and constrained, possibly reflecting structural or contextual limitations on connectivity. Overall, the results indicate that within political groups, connections are relatively rare, but when they do form, people tend to connect with those

who share their views, and these connections are often mutual. However, divisive language is less prevalent, suggesting that controversial discussions might not be as common or influential in shaping these intra-party connections.

Within-cluster analysis shows minimal variance among human coders, with nearly identical estimates for edges (-5.866 to -5.865), party affiliation (0.626 to 0.633), and divisiveness (-0.226 to -0.210), highlighting strong observer consistency. The Combined-coded network is slightly sparser (weights: -12.867), and the Republican-coded network has a marginally stronger divisiveness effect (-0.226). In contrast, GPT-4o diverges with less negative edges (-5.839), higher mutual (0.485 vs. 0.44) and triangle (0.123 vs. 0.113) terms, and a stronger party affiliation effect (0.651), while showing a weaker divisiveness effect (-0.180). These differences indicate GPT-4o infers denser, more reciprocal, and more homophilous networks compared to the robust consensus among human coders.

Table 3: HERGMs' between-cluster estimation

| Coders | Edges | Party Affiliation | Divisiveness | Weights |
|---|---|---|---|---|
| **Democrat** | -5.097*** | 0.659* | 1.607** | -43.531*** |
| **Republican** | -5.061*** | 0.606 | 1.197*** | -47.313*** |
| **Combined** | -5.158*** | 0.456 | 1.520*** | -57.137*** |
| **GPT-4o** | -5.052*** | 0.592 | 1.027** | -49.5924*** |

*Notes: Models included controls for the following structural terms: gwesp (0.5 for Democrat-, Republican-, and Combined-coded tweets; 0.3 for GPT-4o). P-values < 0.1 (*), 0.05 (**), 0.01 (***).*

Table 3 provides the HERGMs' between-cluster estimation. The results show that interactions between opposing political clusters, or groups, are rare and tend to involve more divisive language. The edges' coefficient is highly negative (-5.097 to -5.052, $p < 0.001$), indicating that inter-cluster connections are rare, reinforcing the overall sparsity of the network. The node attribute coefficient for party affiliation (0.659 to 0.592) is positive, but the effect is less pronounced than within-cluster connections, and it is not statistically significant at the conventional levels ($p > 0.05$). This suggests that political homophily plays a weaker role in inter-cluster connections compared to within-cluster ties, although the influence is still present. The divisiveness coefficient is positive and significant (1.607 to 1.027, $p < 0.05$), indicating that more divisive language is associated with a higher likelihood of inter-cluster ties, suggesting that interactions between clusters may be more polarised.

As above, the weights' parameter is consistently large in magnitude and highly negative across all clusters (-57.137 to -43.531, $p < 0.001$), reinforcing the notion of overall network sparsity. Overall, the findings suggest that inter-cluster ties are infrequent and weakly influenced by political homophily, but more likely to involve divisive language, while stronger interactions discourage connections across clusters. In other words, while users tend to connect with others who share their political views, interactions between different political groups often involve more polarising language, reflecting both the divide and attempts to connect across these divides.

Between-cluster analysis reveals greater coder divergence than within clusters. Edges' estimations remain similarly negative (-5.158 to -5.052, $p < 0.001$), reinforcing network sparsity. Human coders often show some variability. The Democrat-coded network shows the strongest party affiliation (0.659, $p < 0.1$) and divisiveness (1.607, $p < 0.01$), while the Combined-coded network shows weaker homophily (0.456, nonsignificant) and stricter sparsity (weights: -57.137 vs. -43.531/-47.313 for Democrats/Republicans). The Republican-coded network falls between, with party affiliation (0.606) and divisiveness (1.197, $p < 0.01$). GPT-4o produces moderate estimates: party affiliation (0.592) aligns closest to the Republican coders, while divisiveness (1.027, $p < 0.01$) is the weakest, suggesting AI underweights cross-cluster polarisation. These contrasts highlight how human coders—particularly Democrats—emphasise partisan and conflict-driven dynamics, while GPT-4o's inferences soften these effects, though all preserve the core pattern of sparse inter-cluster ties.

## 4.6 Robustness checks

### 4.6.1 Multicollinearity checks

There is no evidence of multicollinearity among the two edges' attributes (i.e., weights and levels of polarised language) used in the HERGMs. To do so, I calculate the Pearson correlation coefficient between the two edge attributes in the network: levels of divisiveness and a measure of edge weight, using pairwise complete observations to handle any missing values. Across all sets of coders, the low Pearson correlation coefficients indicate minimal multicollinearity, suggesting that the two variables included in the models are largely independent and unlikely to distort the estimation results.

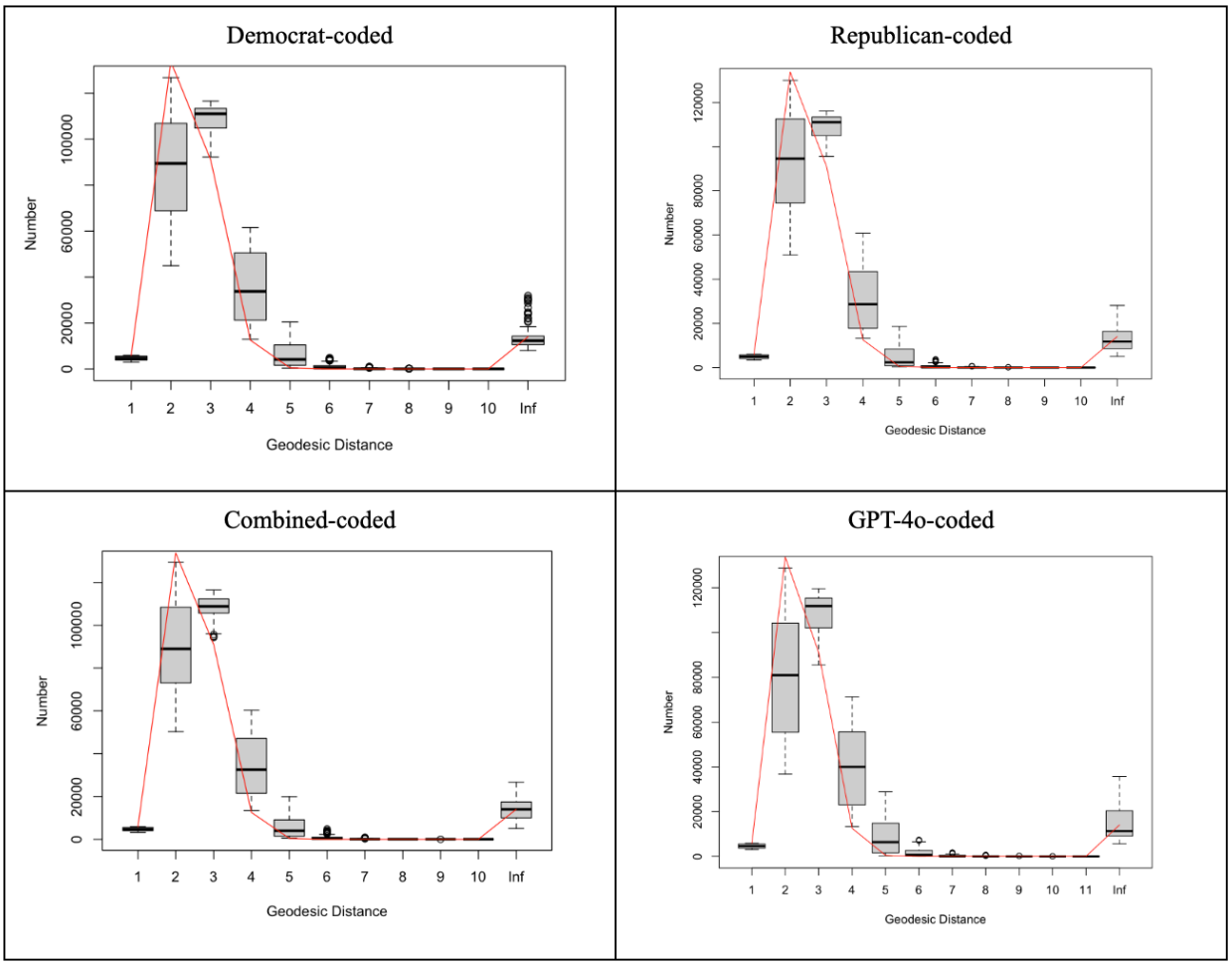
Table 4: Multicollinearity checks

| Coders | Pearson r |
|---|---|
| **Democrat** | 0.0378 |
| **Republican** | 0.0338 |
| **Combined** | 0.0347 |
| **GPT-4o** | 0.0425 |

### 4.6.2 Goodness-of-fit

I now evaluate the HERGMs, via goodness-of-fit assessments, including geodesic distances, in-degree and out-degree distributions as discussed above.

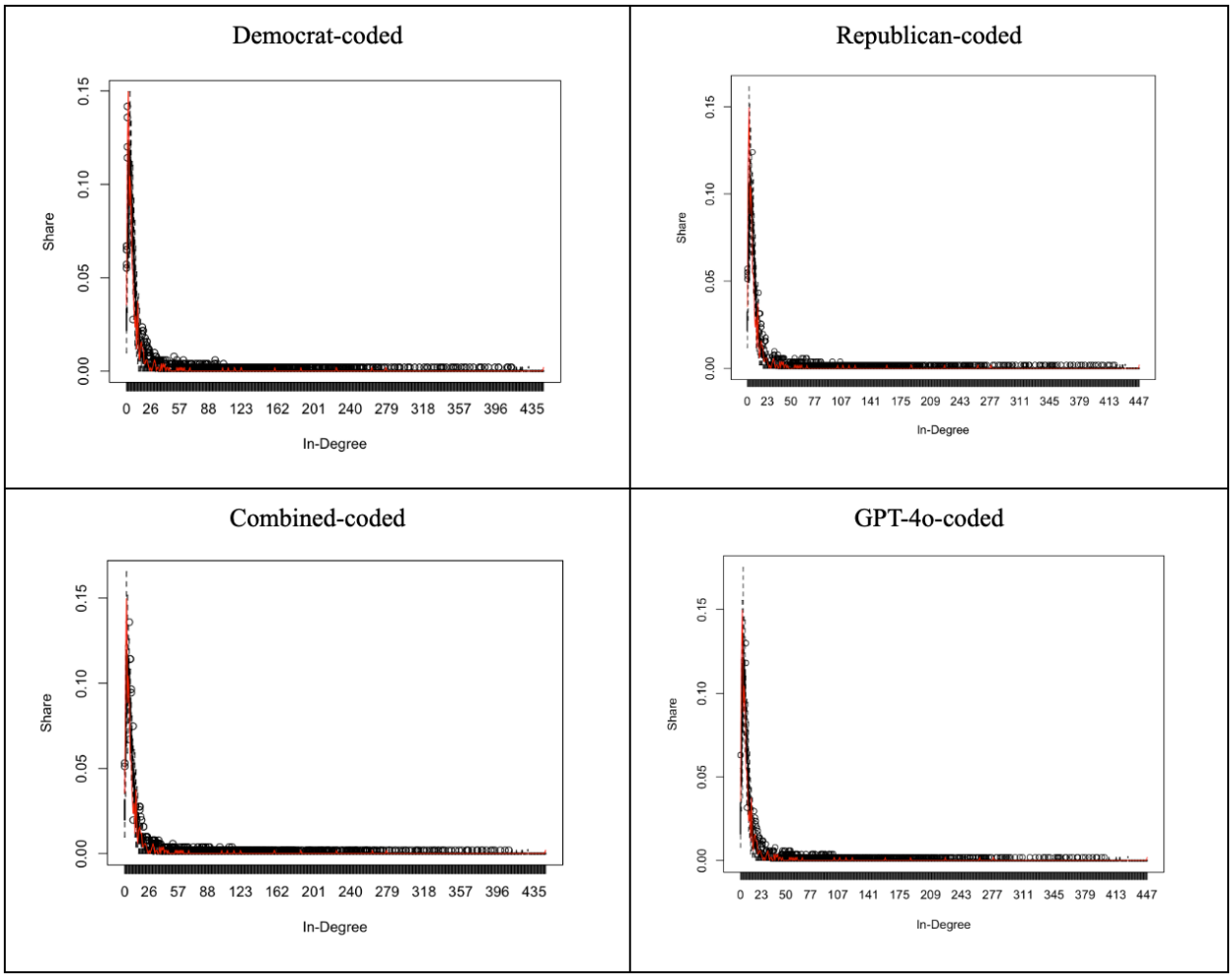
Figure 3: Goodness-of-fit - Geodesic Distance




*Notes: The plot displays the goodness-of-fit assessment for geodesic distance in the HERGM model. The box plots represent the distribution of simulated geodesic distances, while the red line indicates the observed network values. A close alignment between the red line and the box plots suggests a good model fit. The "Inf" mark on the x-axis represents disconnected node pairs.*

Figure 3 displays a goodness-of-fit assessment for geodesic distances in my HERGM, showing that it is a reasonable fit to the observed network structures. The x-axis represents geodesic distances, which measure the shortest path length between pairs of nodes in the network, while the y-axis indicates the number of node pairs at each distance. The box plots summarise the distribution of simulated geodesic distances, with the black horizontal lines representing medians. The red line represents the observed geodesic distance distribution in the empirical network. A good model fit is indicated when the red line falls within the interquartile ranges of the box plots, suggesting that the model-generated networks successfully replicate the observed network's structural properties. In the figure above, the red line follows the simulated distributions well at lower and longer distances, across all models. Thus, overall, the models appear to adequately reproduce the geodesic distance distribution, indicating a reasonable fit to the observed network structures.

Figure 4: Goodness-of-fit - In-degree distribution

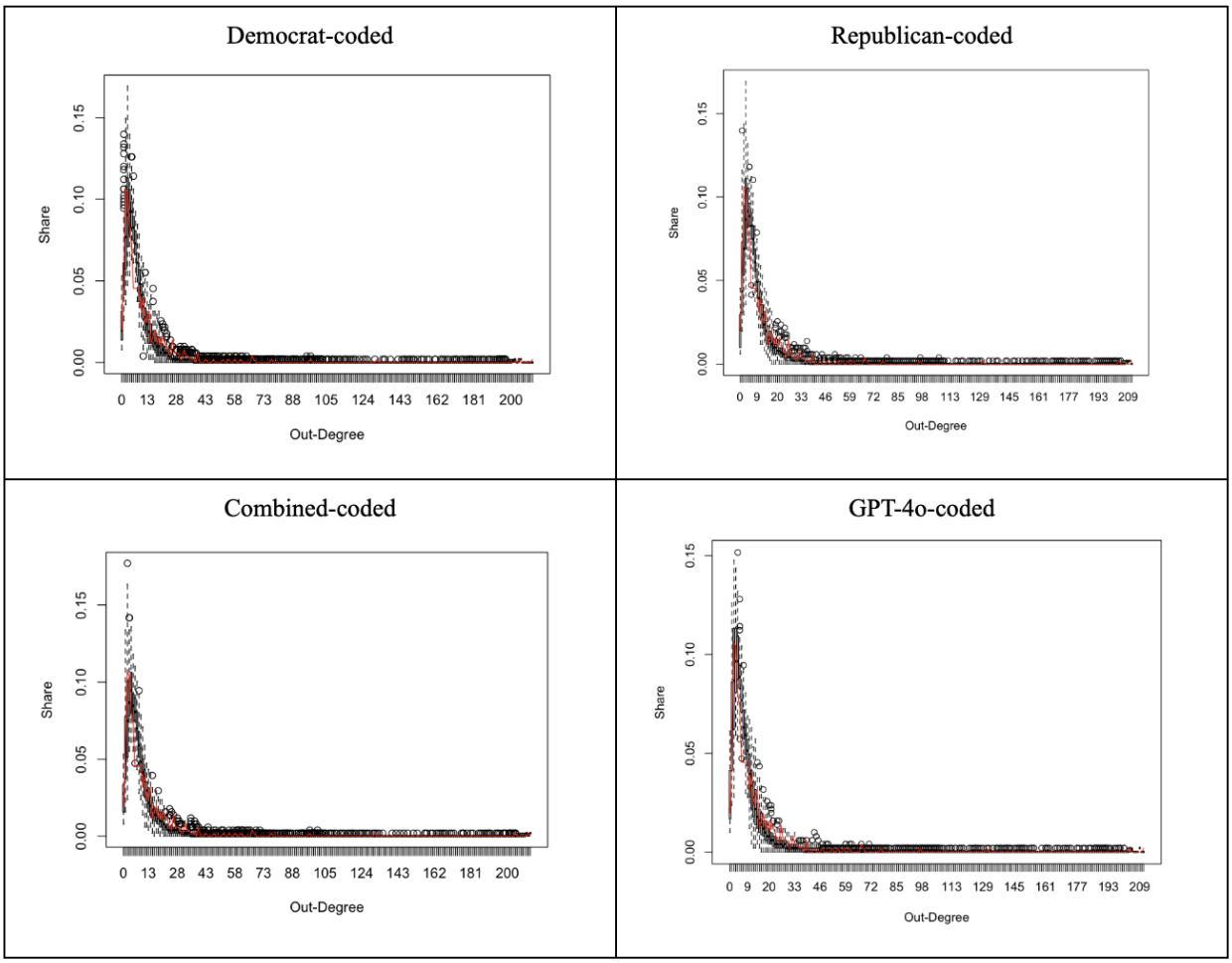



*Notes: This figure displays the in-degree distribution of a network, assessing the goodness-of-fit of the HERGM. The x-axis (In-Degree) represents the number of incoming ties (edges) to nodes. The y-axis (Share) shows the proportion of nodes with a given in-degree. Black dashed lines and points (Simulated Distribution) show the expected in-degree distribution under the fitted ERGM model. The red line (Observed Distribution) represents the actual in-degree distribution in your empirical network. If the red line closely follows the black points, the model accurately captures the observed degree distribution.*

Figure 4 summarises the in-degree distribution and assesses the HERGM's fit, suggesting that the models reliably capture the structural characteristics of the networks. The x-axis represents in-degree (the number of incoming connections to a node), while the y-axis represents the share of nodes with a given in-degree. The black dashed line and points represent the observed in-degree distribution, while the red line represents the simulated distribution generated from the model. The

distribution is highly skewed, with most nodes receiving very few incoming connections, while a small number of nodes exhibit extremely high in-degrees. This suggests a network where a few nodes act as central hubs, receiving a disproportionately high number of connections. Thus, there are a few accounts—likely those of prominent politicians—that attract the majority of attention and engagement within the network. The alignment between the observed and simulated distributions is key to evaluating the model's fit. Here, the red line largely follows the observed data, indicating a reasonable fit, across all models. However, minor deviations, particularly at very low in-degree values, show that the model may slightly mis-estimate the frequency of nodes with minimal connections. Despite this, the models capture the overall structure of the in-degree distributions effectively.

Figure 5: Goodness-of-fit - Out-degree distribution




*Notes:The plot shows the share of nodes with different numbers of outgoing connections. The x-axis represents the out-degree, or the number of ties a node initiates, while the y-axis shows the proportion of nodes with a given out-degree. The black line represent the observed network, while the red line corresponds to the model-simulated distribution. For model assessment, the key aspect is the alignment between the observed and simulated distributions.*

Figure 5 displays the out-degree distribution of the network to assess the goodness-of-fit of the HERGM, showing that the fitted models effectively capture the structure of the networks under analysis. The x-axis represents the out-degree, or the number of outgoing ties initiated by a node, while the y-axis shows the proportion of nodes with a given out-degree. The black box plots represent the observed network distribution, while the red line corresponds to the model-simulated distribution. The distribution is highly right-skewed, with most nodes having very few outgoing connections, while a small number of nodes initiate a disproportionately high number of ties. This

pattern suggests a heterogeneous network, where a few key nodes act as central broadcasters or influential connectors. In other words, a small number of highly active politicians are responsible for initiating most of the outreach and communication within the network. The presence of extreme outliers at the higher end indicates that some nodes are significantly more active than the majority. In terms of model assessment, the red line generally follows the observed distribution, suggesting a reasonable fit, across all models. However, some deviations occur, particularly at lower out-degree values, where the model may slightly underestimate or overestimate node connectivity. Despite these minor discrepancies, the models appear to capture the overall connectivity structure of the networks effectively.

## 4.7 Discussion

This paper investigated how network dynamics and the content of political communication jointly shape polarisation among US politicians on Twitter. Empirically, it combines descriptive centrality measures and hierarchical exponential random graph models (HERGMs) to explore connectivity patterns among Democrats and Republicans. In line with the hypotheses, I find that users' party affiliation strongly predicts mentions (edges) between users from the same party. Moreover, greater levels of divisive language are positively associated with cross-party interactions.

The descriptive findings indicate that while both Democrats and Republicans show similar overall network structures, important differences stand out. Democrats occupy strategically advantageous positions within the political communication network, where they have greater access to and interaction with other highly influential actors. Their statistically greater eigenvector centrality indicates that Democrats are not only well-connected in terms of raw numbers of connections, but are specifically connected to other nodes that themselves have substantial influence and reach. This central network position likely helps Democrats spread their messages more widely and gain greater visibility, giving them a practical advantage in shaping political conversations online. This highlights how being well-connected to influential users can offer strategic benefits that go beyond individual messaging tactics. As such, one can imagine that Democrats may have better capacity to coordinate messaging, respond collectively to emerging issues, and leverage relationships with other influential actors to shape political narratives.

These results contribute to the literature by showing that Democrats and Republicans employ different networking strategies online. Democrats' higher centrality suggests a broader reach and greater influence across the network. These findings are consistent with previous scholarship on the role of centrality in reinforcing ideological echo chambers (Sunstein, 2017; Hobolt et al., 2024; Hong and Sung, 2016), highlighting how network structure can shape communication and mobilisation within each party.

The HERGM results provide deeper insights into network formation. Within political groups, people tend to connect with those who share their views, and these connections are often mutual and contain low levels of polarising language. In this regard, within-cluster analysis shows that intra-cluster ties are relatively sparse, reinforcing the overall sparsity of the network. However, reciprocity is strong, as indicated by the statistically significant mutual term, meaning that if one politician interacts with another, the other is highly likely to reciprocate. Political homophily is also evident, as politicians are significantly more likely to connect with those from the same political group, with party affiliation playing a strong and consistent role in tie formation. Divisiveness slightly discourages within-cluster ties, though its effect is less pronounced than reciprocity and homophily. Moreover, results show that interactions between different political groups often involve more polarising language, reflecting both the divide and attempts to connect across these divides. In other words, mentions across party lines—such as a Democrat mentioning a Republican—are much rarer and more likely to involve polarising or conflictual language, highlighting the divide between political groups and the contentious nature of cross-party engagement.

Consistent with Hypothesis 1, the analysis demonstrates that party affiliation is a strong and consistent predictor of political mentions, reinforcing existing theories of homophily in digital networks (Iyengar et al., 2012; Mason, 2018; Yoon and Park, 2014). Politicians tend to engage within their ideological camp, mirroring prior research showing that online political communication often takes place within like-minded clusters, where co-partisans are both the primary audience and interlocutors (Vergeer, 2020; Esteve Del Valle and Borge Bravo, 2018a). Taken together, these results underscore how digital political communication remains largely confined within partisan boundaries, highlighting the persistent tendency for politicians to engage primarily with their own party and reinforcing the insular nature of online political discourse.

On the contrary, divisive language significantly increases the likelihood of cross-cluster interactions, suggesting that inter-group connections are more conflict-driven than cooperative. This highlights a fragmentation effect, where divisive discourse fosters interactions across clusters rather than shared political affiliation. Furthermore, the negative edges coefficient underscores the limited connectivity across different clusters, while strong intra-cluster interactions reduce the likelihood of inter-cluster ties. The findings that divisive language increases the likelihood of cross-party mentions provides support for Hypothesis 2 and expands on the notion that ideological opponents are not ignored, but rather engaged with—strategically and often antagonistically. This dynamic echoes previous findings from Avalle et al. (2024), Hegelich and Shahrezaye (2015), and Bail et al. (2018), who highlight that cross-party interactions are frequently conflict-driven, marked more by opposition than deliberation. Divisive content appears not to bridge ideological divides constructively, but instead reinforces them by incentivising contentious engagement—what might be considered performative polarisation on a highly visible public platform.

A key contribution of this study is the comparison of coder effects on network estimates. Within-cluster analyses show robust consensus among human coders (e.g., party affiliation and divisiveness scores), with only minor deviations. GPT-4o diverges slightly, inferring denser, more reciprocal networks. Between clusters, human coders—particularly Democrat—emphasise polarisation and party homophily, while GPT-4o's estimates align closer to Republican coding. These discrepancies suggest human coders better capture partisan nuance in contentious interactions, whereas AI may underweight conflict signals. Structural controls (e.g., gwesp parameters) partly explain differences, but methodological choices—human consensus vis-à-vis algorithmic interpretation—critically shape inferred polarisation dynamics. By systematically comparing human and AI coding effects within advanced network models, this study demonstrates how such modelling techniques can more precisely crystallise the underlying structure and dynamics of political communication networks—offering new methodological clarity and helping to bridge gaps in our understanding of how coder perspectives shape interpretations of polarisation and connectivity.

Taken together, the HERGM results emphasise the dominant role of reciprocity, political homophily, and divisiveness in shaping the network's structure. While intra-cluster connections remain sparse, they are reinforced by reciprocity and shared political affiliation. In contrast, inter-cluster ties are infrequent and more likely to emerge in contentious contexts, reinforcing the network's polarisation

and clustering tendencies. These findings underpin the value of combining network analysis with textual analysis in political communication research. While prior studies have often treated network ties and content separately, this approach demonstrates that structural patterns and discursive tone are deeply intertwined, offering a more complete picture of how polarisation unfolds in digital spaces.

This study has some limitations. First, in line with previous literature (Avelle et al., 2024), I use political leaning as a broad proxy for general attitudinal trends, which may oversimplify online discourse. For example, such discourse can be based on further dimensions of leaning, such as news preferences, vaccine attitudes and stance on climate change (Albrecht, 2022; Falkenberg et al., 2022). Secondly, passive users' behaviours remain unaccounted for, meaning divisive language might discourage participation (or users mentioning users from the opposing side), hence generating isolated nodes, which have been excluded in this paper. Nevertheless, the goodness-of-fit (GOF) results suggest that the models presented retain predictive power even in scenarios where isolated nodes are present. While isolated nodes were excluded from the primary analysis due to their lack of mention-based activity, the model's ability to approximate key structural features indicates robustness and generalisability beyond the observed network. In other words, despite their exclusion, the broader dynamics captured by the model may still be reflective of underlying patterns that include more passive or disconnected actors. Third, this study assumes that all mentions are indicative of meaningful interaction. However, mentioning behaviour can be ambiguous. As such, mentions may be used sarcastically, performatively, or simply to amplify content, and not necessarily to engage in genuine political exchange. Deeper qualitative coding or sentiment disambiguation would allow to further analysis of these nuances, even though this paper has provided a strong quantitative measurement.

These limitations suggest new potential directions for future research, particularly in broadening the scope of ideological indicators, incorporating more platforms, in line with Avalle et al. (e.g., Facebook, Telegram and Reddit) and behaviours, and refining both linguistic (e.g., via more substantive qualitative work on the mentions) and network-based measures of political polarisation (e.g., testing alternative modelling strategies, such as stochastic block models or latent space models, which might yield different insights). Future research could include time-dynamics (e.g., pre-, during, post-elections) in line with Barberá et al. (2015), as well as including engagement coefficients (e.g.,

number of likes, retweets, replies) received by the politicians, following Bail's (2022).

# 5 Conclusion

This thesis seeks to advance the theoretical and methodological understanding of political polarisation in US public discourse on Twitter by a multi-level empirical analysis. Across three papers, the research moves from developing and validating new supervised machine learning classifiers for detecting polarising language, to tracing the evolution and sources of polarisation over time, and finally to mapping the network structures that underpin political communication online.

The first chapter contributes a novel measurement framework, introducing four supervised classifiers that annotate text for polarising language along a continuous scale, explicitly accounting for coders' party affiliation and benchmarking against generative AI. The analysis reveals that human coders, especially when their perspectives are combined, consistently rate content as more divisive than AI, with partisan biases evident but not always statistically significant. These findings underscore the importance of integrating multiple perspectives in measurement and highlight both the strengths and limitations of relying on AI's political neutrality, which tends to understate divisiveness. The superior performance of the Combined coders model suggests that aggregating divergent viewpoints can yield more balanced and accurate assessments, while the analysis of coder reliability aligns with existing research on ideological consistency among partisans.

Building on this measurement framework, the second study traces the evolution of polarising language over time and across actor groups. This chapter employs a combination of descriptive time-series analyses, as well as vector-autoregressive models, to study a large-scale dataset of US political tweets. It finds that levels of divisiveness fluctuate in response to major political events, with Republican politicians leading in polarisation during the early period, followed by Democrats and later the public. Media actors consistently exhibit the lowest levels of divisiveness. The dynamic interplay between politicians, the public, and the media is complex: politicians both influence and respond to public sentiment, and cross-party interactions can trigger cycles of escalation and moderation. These findings reveal that polarisation is not static but shaped by reciprocal relationships and contextual events.

The third study shifts focus to the network structure of political discourse, showing that party affiliation is a strong predictor of interaction patterns. To do so, the paper employs hierarchical exponential

random graph models (HERGMs) to predict ties between users based on contextual factors, such as their party affiliation (e.g., Republican or Democrat). Furthermore, I consider the degree of polarised language present in their tweets, as determined by supervised machine learning classifiers based on human coders and artificial intelligence, as developed in paper 1. It finds that party-based homophily dominates network formation. Democrat politicians hold stronger connections to other powerful nodes in the network than Republicans. The most divisive language is concentrated in the cross-party interactions, which are conflict-driven rather than cooperative, reinforcing rather than bridging ideological divides. The integration of network and textual analysis reveals that structural patterns and discursive tone are deeply intertwined, with political communication remaining largely confined within partisan boundaries. Such a complementary approach between network structure and language offers a more complete picture of how polarisation unfolds in digital spaces that is more complete than previous research.

Together, these studies advance our understanding of polarisation as a dynamic process shaped by measurement choices, actor behaviour, and network structure. While theoretical and methodological advances allow for more nuanced detection and analysis, the persistent influence of partisan identity and strategic communication continues to drive fragmentation in online discourse. This work thus provides both new tools for researchers and a theoretical framework of the mechanisms sustaining political polarisation in the digital age.

These findings have significant broader implications. First, they demonstrate that Twitter as a social media platform, while not necessarily being the root cause of polarisation, serves as a powerful facilitator that can amplify partisan divisions and entrench echo chambers. The research shows that exposure to opposing views on social media may not foster greater understanding. In fact, it can sometimes intensify existing attitudes, as users respond defensively or antagonistically to counter-attitudinal content. The architecture of online networks, shaped by both user behaviour and platform algorithms, tends to reinforce ideological homogeneity within groups, while simultaneously accentuating divisions across them. As such, contrary to assumptions about the bridging potential of social media, exposure to opposing views often intensifies affective polarisation. The network analysis reveals that cross-party ties primarily involve adversarial exchanges. This dynamic makes it harder for political exchanges between diverse perspectives to gain traction, thereby impeding opportunities for cooperative deliberation.

Other important implications come from the developed polarisation measurement framework. This framework does not only advance methodological rigour in the study of online polarisation, but also provides evidence for the integration of human judgment and artificial intelligence in social science research. By accounting for coder backgrounds and biases, it addresses longstanding concerns about the subjectivity and replicability of content analysis for more transparent and accountable measurement practices. The approach shows that leveraging diverse perspectives, rather than defaulting to the supposed objectivity of humans and AI, can lead to richer and more nuanced insights into contentious discourse. More broadly, this work highlights the need for interdisciplinary collaboration in developing tools that can adapt to evolving language and political contexts, ensuring that future analyses remain robust as both technology and public debate continue to change.

These insights underscore the need for interventions that go beyond simply increasing exposure to opposing viewpoints. Effective strategies may include promoting digital and media literacy, designing algorithms that encourage more diverse and balanced content exposure, and fostering environments that reward constructive engagement over online forms of polarisation. The methodological advances presented here also offer valuable tools for researchers, policymakers, and platform designers seeking to monitor and mitigate the risks of polarisation in digital spaces. Ultimately, the combination of these chapters demonstrates that understanding and addressing political polarisation requires a multi-layered approach—one that accounts for individual perceptions, temporal behaviour, and the network structures of online communication.

Although each paper has addressed its specific limitations, it is important to recognise that this thesis as a whole is also subject to certain constraints that warrant acknowledgment. First, by exclusively focusing on Twitter, the analysis may not fully capture the spectrum of political discourse in the US. It is plausible that significant conversations also occur on other social media platforms (e.g., Facebook, Bluesky, etc.), in traditional media (e.g., interviews), and via press-releases. Additionally, the platform under analysis, Twitter, has undergone significant changes in ownership, policies, and user demographics, and algorithmic behaviour. In particular, several changes have occurred after 2022 when Elon Musk, who has since become a controversial figure in American politics, acquired the platform. These changes may limit the generalisability of the findings over time, to more recent or future contexts. Another threat to external validity is presented by the highly polarising political context, party system, and the English language used in the US. Specifically, the textual

measurement approach developed in Paper 1, which then served the remaining chapters, may not be directly applicable to multiparty systems, emerging democracies, or different languages. For example, in European countries, party systems are often more fragmented and citizens tend to show weaker partisan attachments (Huddy et al., 2018; 2020; Kriesi et al., 2006; 2012). In Latin America, polarisation is frequently rooted in the personalisation of politics, with charismatic or populist leaders and movements playing a central role (Areal, 2022; Torcal and Carty, 2023; Torcal and Harteveld, 2024). In newer democracies, where institutional trust and party systems may not be fully established, polarisation can be driven more by ethnic, religious, or regional divisions than by established partisan identities (Reiljan et al., 2024; Torcal, 2023; Torcal and Harteveld, 2024). Additionally, even among English-speaking countries such as the UK or Australia, linguistic and cultural differences may affect how polarising language is used and interpreted, potentially limiting the generalisability of these results. Furthermore, the detection of polarised language relies on supervised machine learning classifiers, which, while robust, may miss subtle rhetorical strategies such as sarcasm or coded language. The temporal scope of the analysis, limited to the years 2015–2020, also means that shifts in communication patterns outside this window are not addressed.

Despite these constraints, the thesis demonstrates the value of combining methodological innovation with multi-level analysis to deepen our understanding of polarisation in digital public spheres. The integration of measurement, temporal, and network approaches provides a flexible framework that can be adapted and extended to new contexts, languages, and platforms—provided that future research carefully validates and calibrates these tools to account for new political, linguistics, and institutional contexts. Ultimately, this work offers both new insights and practical tools for researchers, policymakers, and platform designers seeking to understand and address the challenges of political polarisation in the digital age.

# Bibliography


Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, *70*(2), 542–555.

Albrecht, D. (2022). Vaccination, politics and COVID-19 impacts. *BMC Public Health*, *22*, 96.

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M., & Ladwig, P. (2013). The 'Nasty Effect': Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*, 373–387.

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*, 373–387.

Ansolabehere, S., & Iyengar, S. (1995). *Going negative: How attack ads shrink and polarize the electorate*. The Free Press.

Appel, R., Pan, J., & Roberts, M. E. (2023). Partisan conflict over content moderation is more than disagreement about facts. *Science Advances*, *9*(44), eadg6799.

Areal, J. (2022). 'Them' without 'us': Negative identities and affective polarization in Brazil. *Political Research Exchange*, *4*(1).

Aref, S., & Neal, Z. P. (2020). Detecting coalitions by optimally partitioning signed networks of political collaboration. *Scientific Reports*, *10*(1), 1506.

Aref, S., & Neal, Z. P. (2021). Identifying hidden coalitions in the U.S. House of Representatives by optimally partitioning signed networks based on generalized balance. *Scientific Reports*, *11*(1), 1–9.

Arnaboldi, V., Passarella, A., Conti, M., & Dunbar, R. (2017). Structure of ego-alter relationships of politicians in Twitter. *Journal of Computer-Mediated Communication*, *22*(5), 231–247.

Ashokkumar, A., Talaifar, S., Fraser, W. T., Landabur, R., Buhrmester, M., Gómez, Á., Paredes, B., & Swann, W. B. Jr. (2020). Censoring political opposition online: Who does it and why. *Journal of Experimental Social Psychology*, *91*, 104031.

Asimovic, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences*, *118*(25), e2022819118.

Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Information, Communication & Society*, *16*(3), 291–314.

Auter, Z. J., & Fine, J. A. (2016). Negative campaigning in the social media age: Attack advertising on Facebook. *Political Behavior*, *38*, 999–1020.

Avalle, M., Di Marco, N., Etta, G., et al. (2024). Persistent interaction patterns across social media platforms and over time. *Nature*, *628*, 582–589.

Bail, C. (2021). *Breaking the Social Media Prism*. Princeton: Princeton University Press.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, M., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Bail, C. A. (2015). *Terrified: How anti-Muslim fringe organizations became mainstream*. Princeton University Press.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*, 1130–1132.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91.

Barberá, P., Casas, A., Nagler, J., Egan, P., Bonneau, R., Jost, J., & Tucker, J. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, *113*(4), 883–901.

Barberá, P., Gohdes, A., Zeitoff, T., & Iakhins, E. (2019). From riot police to tweets: How world leaders use social media during contentious politics. *Working paper*.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Barberá, P. (2020). Social media, echo chambers, and political polarisation. In N. Persily & J. Tucker (Eds.), *Social Media and Democracy: The State of the Field*. Cambridge: Cambridge University Press.

Bawn, K., Cohen, M., Karol, D., Masket, S., Noel, H., & Zaller, J. (2012). A theory of political parties: Groups, policy demands and nominations in American politics. *Perspectives on Politics*, *10*(3), 571–597.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowdsourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.

Bode, L., & Vraga, E. K. (2018). Studying politics across media. *Political Communication*, *35*(1), 1–7.

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, 28(4), 466–484.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). Analyzing social networks. London: Sage.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895.

Boutet, A., Kim, H., & Yoneki, E. (2012). What's in Twitter: I know what parties are popular and who you are supporting now! In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining* (pp. 129–132).

Box-Steffensmeier, J. M., Freeman, J. R., Hitt, M. P., & Pevehouse, J. C. W. (2014). *Time-series Analysis for the Social Sciences*. Cambridge: Cambridge University Press.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS Nexus*, *114*(28), 7313–7318.

Brandt, P. T., & Williams, J. T. (2007). *Multiple time-series Models*. Thousand Oaks, CA: SAGE.

Bright, J. (2018). Explaining the emergence of echo chambers on social media: the role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*, 17–33.

Bright, J. (2018). Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of Computer-Mediated Communication*, *32*(1), 17–33.

Budhiraja, A., & Pal, J. (2020). Twitter and political culture: Short text embeddings as a window into political fragmentation. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 335–336.

Campbell, J. E. (1982). Cosponsoring legislation in the U.S. Congress. *Legislative Studies Quarterly*, *7*(3), 415–422.

Capozzi, A., Semeraro, A., & Ruffo, G. (2023). Analyzing and visualizing polarization and balance with signed networks: The U.S. Congress case study. *Journal of Complex Networks*, *11*(4), cnad027.

Carlson, D., & Montgomery, J. M. (2017). A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts. *American Political Science Review*, *111*(4), 835–843.

Castanho Silva, B., & Proksch, S.-O. (2022). Politicians unleashed? Political communication on Twitter and in parliament in Western Europe. *Political Science Research and Methods*, 10(4), 776–792.

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020). COVID-19 Government Response Event Dataset (CoronaNet v.1.0). *Nature Human Behaviour*, *4*, 756–768.

Cherepnalkoski, D., Karpf, A., Mozetič, I., & Grčar, M. (2016). Cohesion and coalition formation in the European Parliament: Roll-call votes and Twitter activities. *PloS One*, *11*(11), e0166586.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. *ICWSM*, *133*, 89–96.

Conover, M. D., Goncalves, B., Flammini, A., & Menczer, F. (2012). Partisan Asymmetries in Online Political Activity. *EPJ Data Science*, *1*(1), 1–19.

Costa, M., Briggs, T., Chahal, A., Fried, J., Garg, R., Kriz, S., Lei, L., Milne, A., & Slayton, J. (2020). How partisanship and sexism influence voters' reactions to political #MeToo scandals. *Research & Politics*, *7*(3).

Crabtree, C., Golder, M., Gschwend, T., & Indriason, I. H. (2020). It is not only what you say, it is also how you say it: The strategic use of campaign sentiment. *The Journal of Politics*, *82*(3), 1044–1060.

de Jong, J. F. (2024). Cross-partisan discussions reduced political polarization between UK voters, but less so when they disagreed. *Communications Psychology*, *2*(5).

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on Facebook. *Scientific Reports*, *6*(1), 37825.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN, USA.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

De Vries, C. E., Bakker, B. N., Hobolt, S. B., & Arceneaux, K. (2021). Crisis signaling: How Italy's coronavirus lockdown affected incumbent support in other European countries. *Political Science Research and Methods*, *9*(3), 451–467.

de Zúñiga, H. G., & Valenzuela, S. (2011). The mediating path to a stronger citizenship: Online and offline networks, weak ties, and civic engagement. *Communication Research*, *38*(3), 397–421.

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745.

Duggan, M., & Smith, A. (2016). The political environment on social media. *Pew Research Center*.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29.

Egan, P. J. (2013). *Partisan Priorities: How Issue Ownership Drives and Distorts American Politics*. New York: Cambridge University Press.

Enli, G., & Skogerbø, E. (2013). Personalized campaigns in party-centered politics: Twitter and Facebook as arenas for political communication. *Information, Communication & Society*, *16*(5), 757–774.

Esteve Del Valle, M., & Borge Bravo, R. (2018). Echo chambers in parliamentary Twitter networks: The Catalan case. *International Journal of Communication*, *12*, 1715–1735.

Esteve Del Valle, M., Broersma, M., & Ponsioen, A. (2022). Political interaction beyond party lines: Communication ties and party polarization in parliamentary twitter networks. *Social Science Computer Review*, *40*(3), 736–755.

Esteve Del Valle, M., Sijtsma, R., Stegeman, H., & Borge, R. (2020). Online deliberation and the public sphere: Developing a coding manual to assess deliberation in twitter political networks. *Javnost—The Public*, *27*(3), 211–229.

Evans, H. K., Cordova, V., & Sipole, S. (2014). Twitter style: An analysis of how house candidates used Twitter in their 2012 campaigns. *PS: Political Science & Politics*, *47*(2), 454–462.

Falkenberg, M., et al. (2022). Growing polarization around climate change on social media. *Nature Climate Change*, *12*, 1114–1121.

Finkel, S. E., & Geer, J. G. (1998). A spot check: Casting doubt on the demobilizing effect of attack advertising. *American Journal of Political Science*, *42*(2), 573-595.

Flamino, J., Galeazzi, A., Feldman, S., et al. (2023). Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nature Human Behaviour*, *7*, 904-916.

Fletcher, R., Cornia, A., & Nielsen, R. K. (2020). How Polarized Are Online and Offline News Audiences? A Comparative Analysis of Twelve Countries. *The International Journal of Press/Politics*,

*25*(2), 169-195.

Fontan, A., & Altafini, C. (2021). A signed network perspective on the government formation process in parliamentary democracies. *Scientific Reports*, *11*, 5134.

Fowler, J. H. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis*, *14*(4), 456-487.

Fraxanet, E., Pellert, M., Schweighofer, S., Gómez, V., & Garcia, D. (2024). Unpacking polarization: Antagonism and alignment in signed networks of online interaction. *PNAS Nexus*, *3*(12), pgae276.

Friedrichs, J., Stoehr, N., & Formisano, G. (2022). Fear-anger contests: Governmental and populist politics of emotion. *Online Social Networks and Media*, *32*, 100240.

Friedrichs, J., Stoehr, N., & Formisano, G. (2022). Fear-anger cycles: Governmental and populist politics of emotion. *Online Social Networks and Media*, *32*, 100240.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-directed Placement. *Software - Practice and Experience*, *21*(11), 1129-1164.

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference* (pp. 913-922).

Garrett, R. K., Gvirsman, S. D., Johnson, B. K., Tsfati, Y., Neo, R., & Dal, A. (2014). Implications of pro- and counter-attitudinal information exposure for affective polarization. *Human Communication Research*, *40*(3), 309-332.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, *14*(2), 265-285.

Gerber, E. R., & Morton, R. B. (1998). Primary Election Systems and Representation. *Journal of Law, Economics, and Organization*, *14*(2), 304-324.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms CrowdWorkers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, *120*(3), e2305016120.

Goi, S. (2005). Agonism, Deliberation, and the Politics of Abortion. *Polity*, *37*(1), 54-81.

Gonzalez-Bailón, S., & Wang, N. (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks*, *44*, 95-104.

Gould, R. V., & Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, 89-126.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, *91*(3), 481-510.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*.

Gross, J. H., & Johnson, K. T. (2016). Twitter taunts and tirades: Negative campaigning in the age of Trump. *PS: Political Science & Politics*, *49*(4), 748-754.

Gruzd, A., & Roy, J. (2014). Investigating political polarization on Twitter: A Canadian perspective. *Policy & Internet*, *6*(1), 28-45.

Gruzd, A., & Tsyganova, A. (2015). Information wars and online activism during the 2013/2014 crisis in Ukraine: Examining the social structures of pro- and anti-Maidan groups. *Policy & Internet*, *7*(2), 121-158.

Guerra, P., Meira Jr., W., Cardie, C., & Kleinberg, R. (2021). A measure of polarization on social media networks based on community boundaries. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 215-224.

Guess, A. M., et al. (2023a). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*, 398-404.

Guess, A. M., et al. (2023b). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, *381*, 404-408.

Guess, M. A., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Moehler, D., Nyhan, B., Valesco Rivera, C., Settle, J., Thomas, D. R., Thorson, E., & Tucker, J. A. (2023a). How do social media feed algorithms affect attitudes and behaviour in an election campaign?

*Science*, *381*(6656), 398-404.

Guess, M. A., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Moehler, D., Nyhan, B., Valesco Rivera, C., Settle, J., Thomas, D. R., Thorson, E., & Tucker, J. A. (2023b). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, *381*(6656), 404-408.

Hall, R. L. (1992). Measuring legislative influence. *Legislative Studies Quarterly*, 205-231.

Hare, C., & Poole, K. T. (2014). The polarization of contemporary American politics. *Polity*, *46*(3), 411-429.

Harteveld, E., Berntzen, L. E., Kokkonen, A., Kelsall, H., Linde, J., & Dahlberg, S. (2022). The (alleged) consequences of affective polarization: A survey experiment in 9 countries. Retrieved from `https://osf.io/preprints/osf/64uwd` (Accessed: 6 November 2024).

Harteveld, E. (2021). Fragmented foes: Affective polarization in the multiparty context of the Netherlands. *Electoral Studies*, *71*, 102332.

Hawkins, K. A., Rovira Kaltwasser, C., & Andreadis, I. (2020). The activation of populist attitudes. *Government and Opposition*, *55*(2), 283-307.

Hegelich, S., & Shahrezaye, M. (2015). The communication behavior of German MPs on Twitter: Preaching to the converted and attacking opponents. *European Policy Analysis*, *1*(2), 155-174.

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analysis to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 154-174.

Hobolt, S., Leeper, T., & Tilley, J. (2020). Divided by the vote: Affective polarization in the wake of Brexit. *British Journal of Political Science*, 1-18.

Hobolt, S. B., Lawall, K., & Tilley, J. (2024). The polarizing effect of partisan echo chambers. *American Political Science Review*, *118*(3), 1-16.

Holbert, R. L., Garrett, R. K., & Gleason, L. S. (2010). A new era of minimal effects? A response to Bennett and Iyengar. *Journal of Communication*, *60*(1), 15-34.

Holbrook, T. M., & McClurg, S. D. (2005). The mobilization of core supporters: Campaigns, turnout, and electoral composition in the United States presidential elections. *American Journal of Political Science*, *49*(4), 689-703.

Holland, P. W., & Leinhardt, S. (1972). Holland and Leinhardt reply: Some evidence on the transitivity of positive interpersonal sentiment. *American Journal of Sociology*, *77*(6), 1205-1209.

Hong, H., & Sung, K. (2016). Political polarization on Twitter: Implication for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777-782.

Hong, S., & Kim, S. H. (2016). Political polarization on Twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*, 777-782.

Hopp, T., & Vargo, C. J. (2017). Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data. *Computers in Human Behavior*, *68*, 368-377.

Huang, H.-Y., Wu, Z., Yang, Y., Zhang, J., & Wu, Y. (2025). Unlocking the power of LLM uncertainty for active in-context example selection. *arXiv:2408.09172*.

Huckfeldt, R., Johnson, P. E., & Sprague, J. (2004). *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. Cambridge: Cambridge University Press.

Huddy, L., Bankert, A., & Davies, C. (2018). Expressive versus instrumental partisanship in multiparty European systems. *Political Psychology*, *39*, 173-199.

Huddy, L., Davies, C., & Sandor, J. (2020). Measuring the direction and strength of partisan identity. In Oscarsson, H., & Holmberg, S. (Eds.), *Research Handbook on Political Partisanship*. Edward Elgar.

Huddy, L., Sears, D. O., Levy, J. S., & Jerit, J. (2023). *The Oxford Handbook of Political Psychology*. Oxford University Press.

Hughes, A., & Wojcik, S. (2019). 10 facts about Americans and Twitter. Pew Research Center. Retrieved from `https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/` (Accessed: 10 April 2023).

Intal, C., & Yasseri, T. (2021). Dissent and rebellion in the House of Commons: A social network

analysis of Brexit-related divisions in the 57th Parliament. *Applied Network Science*, *6*, 36.

Iyengar, S., & Wagner, W. (2024). Conceptual debates and approaches. In Torcal, M., & Harteveld, E. (Eds.), *Handbook of Affective Polarization*. Retrieved from `https://osf.io/86ds3/` (Accessed: 6 November 2024).

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarisation. *Public Opinion Quarterly*, *76*(3), 405-431.

Iyengar, S., & Westwood, S. J. (2014). Fear and loathing across party lines: New evidence on group polarisation. *American Journal of Political Science*, *59*(3), 690-707.

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford and New York: Oxford University Press.

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, *13*(1), 72-91.

Kalmoe, N. P. (2014). Fueling the fire: Violent metaphors, trait aggression, and support for political violence. *Political Communication*, *31*(4), 545-563.

Kastellec, J. P., Lax, J. R., Malecki, M., & Phillips, J. H. (2015). Polarizing the electoral connection: Partisan representation in Supreme Court confirmation politics. *The Journal of Politics*, *77*(3), 787-804.

Kessler, D., & Krehbiel, K. (1996). Dynamics of cosponsorship. *American Political Science Review*, *90*(3), 555-566.

Kiley, J. (2017). In polarized era, fewer Americans hold a mix of conservative and liberal views. Pew Research Center. Retrieved from `https://www.pewresearch.org/politics/2017/10/05/in-polarized-era-fewer-americans-hold-a-mix-of-conservative-and-liberal-views/` (Accessed: 3 March 2024).

Knobloch-Westerwick, S., & Kleinman, S. (2012). Pre-election selective exposure: Confirmation bias versus informational utility. *Communication Research*, *39*(2), 170-193.

Kosmidis, S., Hobolt, S. B., Molloy, E., & Whitefield, S. (2019). Party competition and emotive rhetoric. *Comparative Political Studies*, *52*(6), 811-837.

Kosmidis, S., & Theocharis, I. (2020). Can social media incivility induce enthusiasm? Evidence from survey experiments. *Public Opinion Quarterly*, *84*(1), 284-308.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), e2210666120.

Krehbiel, K. (1995). Cosponsors and wafflers from A to Z. *American Journal of Political Science*, 906-923.

Kriesi, H., Grande, E., Dolezal, M., Helbling, M., Höglinger, D., Hutter, S., & Wüest, B. (2012). *Political conflict in Western Europe*. Cambridge University Press.

Kriesi, H., Grande, E., Lachat, R., Dolezal, M., Bornschier, S., & Frey, T. (2006). Globalization and the transformation of the national political space: Six European countries compared. *European Journal of Political Research*, *45*(6), 921-956.

Kubin, E., & Von Sikorski, C. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, *45*(3), 188-206.

Kühne, R., & Schemer, C. (2015). The emotional effects of news frames on information processing and opinion formation. *Communication Research*, *42*(3), 387-407.

Kühne, R., Schemer, C., Matthes, J., & Wirth, W. (2011). Affective priming in political campaigns: How campaign-induced emotions prime political opinions. *International Journal of Public Opinion Research*, *23*(4), 485-507.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Lasco, G., & Curato, N. (2019). Medical populism. *Social Science & Medicine*, *221*, 1-8.

Lau, R. R., Sigelman, L., & Rovner, I. B. (2007). The effects of negative political campaigns: A meta-analytic reassessment. *The Journal of Politics*, *69*(4), 1176-1209.

Lawrence, R. G., & Boydstun, A. E. (2017). What I should really be asking about media attention to Trump. *Political Communication*, *34*(1), 150-153.

Layman, G. C., Carsey, T. M., & Horowitz, J. M. (2006). Party polarization in American politics:

Characteristics, causes, and consequences. *Annual Political Science Review*, *9*, 83-110.

Lelkes, Y., & Sniderman, P. M. (2017). The ideological asymmetry of the American party system. *British Journal of Political Science*, *46*(4), 825-844.

Levendusky, M. S. (2013). Partisan media exposure and attitudes toward the opposition. *Political Communication*, *30*(4), 565-581.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831-870.

Liang, H., & Zhang, X. (2021). Partisan bias of perceived incivility and its political consequences: Evidence from survey experiments in Hong Kong. *Journal of Communication*, *71*(3), 357-379.

Lin, H., Wang, Y., Lee, J., & Kim, Y. (2023). The effects of disagreement and unfriending on political polarization: A moderated-mediation model of cross-cutting discussion on affective polarization via unfriending contingent upon exposure to incivility. *Journal of Computer-Mediated Communication*, *28*(4), 1-13.

Lindner, N. M., & Nosek, B. A. (2009). Alienable speech: Ideological variations in the application of free-speech principles. *Political Psychology*, *30*(1), 67-92.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Lütkepohl, H. (2005). *New Introduction to Multiple Time-Series Analysis*. Springer.

Lütkepohl, H. (2008). Impulse response function. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

MacKuen, M., Wolak, J., Keele, L., & Marcus, G. E. (2010). Civic engagements: Resolute

partisanship or reflective deliberation. *American Journal of Political Science*, *54*(2), 440-458.

Marozzo, F., & Bessi, A. (2018). Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, *8*, 1-13.

Martin, P. S. (2004). Inside the black box of negative campaign effects: Three reasons why negative campaigns mobilize. *Political Psychology*, *25*(4), 545-562.

Mason, L. (2013). The rise of uncivil agreement: Issue versus behavioural polarization in the American electorate. *American Behavioral Scientist*, *57*, 140-159.

Mason, L. (2015). I disrespectfully agree: The differential effects of partisan sorting on social and issue polarisation. *American Journal of Political Science*, *59*(1), 128-145.

Mason, L. (2018). Ideologues without issues: The polarizing consequences of ideological identities. *Public Opinion Quarterly*, *82*(S1), 280-301.

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*, 375-441.

Mayhew, D. R. (2004). *Congress: The Electoral Connection*. Yale University Press.

McLoughlin, L., & Ward, S. (2020). Turds, traitors and tossers: The abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, *26*(1), 47-73.

Monge, P. R., & Contractor, N. S. (2003). *Theories of communication networks*. Oxford University Press.

Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, *68*(1), 103-127.

Mouffe, C. (2000). *The Democratic Paradox*. Verso.

Müller, S. (2022). The temporal focus of campaign communication. *The Journal of Politics*, *84*(1), 585-590.

Mutz, D. C., & Mondak, J. J. (2006). The workplace as a context for cross-cutting political discourse. *Journal of Politics*, *68*(1), 140-155.

Mutz, D. C. (2002). Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, *96*(1), 111-126.

Nai, A., & Maier, J. (2021). Is negative campaigning a matter of taste? Political attacks, incivility, and the moderating role of individual differences. *American Politics Research*, *49*(3), 269-281.

Nai, A., & Maier, J. (2023). Mediatized campaign attacks fuel affective polarization if perceived as negative: Experimental evidence with American voters. *International Journal of Communication*, *17*, 5246-5267.

Nai, A., & Young, E. L. (2024). Negative campaigning and affective polarization. In Torcal, M., & Harteveld, E. (Eds.), *Handbook of Affective Polarization*. Retrieved from `https://osf.io/86ds3/` (Accessed: 6 November 2024).

Németh, R. (2023). A scoping review on the use of natural language processing in research on political polarization: Trends and research prospects. *Journal of Computational Social Science*, *6*(1), 289-313.

Neundorf, A., & Adams, J. (2018). The micro-foundations of party competition and issue ownership: The reciprocal effects of citizens' issue salience and party attachments. *British Journal of Political Science*, *48*(2), 385-406.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2017). *Reuters Institute Digital News Report 2017*. Retrieved from `https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web-0.pdf`

Nguyen, D. Q., Vu, T., & Tuan, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. *arXiv:2005.10200*.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Ogawa, Y., Miyata, K., Yamamoto, H., & Ikeda, K. (2013). Association between selective exposure and attitude on Twitter. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (Vol. 3, pp. 122–125).

Panning, W. H. (1982). Blockmodels: From relations to configurations. *American Journal of*

*Political Science*, 585–608.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Pellegrini, P. A., & Grant, J. T. (1999). Policy coalitions in the US Congress: A spatial duration modeling approach. *Geographical Analysis*, *31*(1), 45–66.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Peralta, A. F., Ramaciotti, P., Kertész, J., & Iñiguez, G. (2024). Multidimensional political polarization in online social networks. *Physical Review Research*, *6*(1), 013170.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751.

Pew Research Center. (2022). *As partisan hostility grows, signs of frustration with the two-party system*. Retrieved from https://www.pewresearch.org/politics/2022/08/09/as-partisan-hostility-grows-signs-of-frustration-with-the-two-party-system/

Pew Research Center. (2024). *Social Media Fact Sheet*. Retrieved from https://www.pewresearch.org/internet/fact-sheet/social-media/

Pew Research Center. (2023). *Americans' dismal views of the nation's politics*. Retrieved from https://www.pewresearch.org/politics/2023/09/19/americans-dismal-views-of-the-nations-politics/

Qiu, J., Lin, Z., & Shuai, Q. (2019). Investigating the opinions distribution in the controversy on social media. *Information Sciences*, *489*, 274–288.

Quelle, D., & Bovet, A. (2025). Bluesky: Network topology, polarization, and algorithmic curation. *PLoS ONE*, *20*(2), e0318034.

Rachev, S. T., Mittnik, S., Fabozzi, F. J., Focardi, S. M., & Jasic, T. (2021). *Financial econometrics: From basics to advanced modelling techniques*. John Wiley & Sons.

Reiljan, A., Garzia, D., Ferreira Da Silva, F., & Trechsel, A. H. (2024). Patterns of affective polarization toward parties and leaders across the democratic world. *American Political Science Review*, *118*(2), 654–670.

Reisman, G., & Dvir-Gvirsman, S. (2024). Social media and affective polarization. In M. Torcal & E. Harteveld (Eds.), *Handbook of Affective Polarization*. Retrieved from `https://osf.io/86ds3/`

Reiter, F., & Matthes, J. (2022). "The good, the bad, and the ugly": A panel study on the reciprocal effects of negative, dirty, and positive campaigning on political distrust. *Mass Communication and Society*, *25*(5), 649–672.

Rodriguez, P., & Spirling, A. (2021). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *Journal of Politics* (Forthcoming).

Röllicke, L. (2023). Polarisation, identity and affect: Conceptualising affective polarisation in multi-party systems. *Electoral Studies*, *85*, 102655.

Ross Arguedas, A., Robertson, C., Fletcher, R., & Nielsen, R. (2022). *Echo chambers, filter bubbles, and polarisation: A literature review*. Reuters Institute for the Study of Journalism.

Schiller, W. J. (1995). Senators as political entrepreneurs: Using bill sponsorship to shape legislative agendas. *American Journal of Political Science*, 186-203.

Schöll, N., Gaël, G., & Le Mens, G. (2021). Politician-citizen interactions and dynamic representation: Evidence from Twitter. *Barcelona GSE Working Paper Series*.

Sharber, B. (2020). *Analyzing political polarization in news media with natural language processing* (Doctoral dissertation, Middle Tennessee State University).

Simchon, A., Brady, W. J., & Van Bavel, J. J. (2022). Troll and divide: The language of online polarization. *PNAS Nexus*, 1(1).

Song, H. (2015). Uncovering the structural underpinnings of political discussion networks: Evidence from an exponential random graph model. *Journal of Communication*, 65(1), 146-169.

Stavrakakis, Y. (2018). Paradoxes of polarization: Democracy's inherent division and the (anti-) populist challenge. *American Behavioral Scientist*, 62(1), 43-58.

Stevens, D. (2008). The relationship between negative political advertising and public mood: Effects and consequences. *Journal of Elections, Public Opinion and Parties*, 18(2), 153-177.

Stromer-Galley, J. (2000). Online interaction and why candidates avoid it. *Journal of Communication*, 50(4), 111-132.

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Sydnor, E. (2019). *Disrespectful democracy: The psychology of political incivility*. Columbia University Press.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755-769.

Talbert, J. C., & Potoski, M. (2002). Setting the legislative agenda: The dimensional structure of bill cosponsoring and floor voting. *Journal of Politics*, 64(3), 864-891.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open*, 10(2), 1-15.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, 66(6), 1007-1031.

Tian, T., Zhu, J., & Qiaoben, Y. (2018). Max-margin majority voting for learning from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10), 2480-2494.

Torcal, M., & Carty, E. (2022). Partisan sentiments and political trust: A longitudinal study of Spain. *South European Society and Politics*, 27(1), 171-196.

Torcal, M., & Harteveld, E. (2024). Introduction. In M. Torcal & E. Harteveld (Eds.), *Handbook*

*of Affective Polarization*. Retrieved from `https://osf.io/86ds3/`

Torcal, M. (2023). *Affective polarisation in Spain: Electoral, regional and media conflictuality*. Taylor & Francis.

Törnberg, P. (2024). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 0(0), 1-15.

Tromble, R. (2018). Thanks for (actually) responding! How citizen demand shapes politicians' interactive practices on Twitter. *New Media & Society*, 20(2), 676-697.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*.

Tucker, J. A., Theocharis, Y., Roberts, M. E., & Barberá, P. (2017). From liberation to turmoil: Social media and democracy. *Journal of Democracy*, 28(4), 46-59.

Valentim, V., & Widmann, T. (2021). Does radical-right success make the political debate more negative? Evidence from emotional rhetoric in German state parliaments. *Political Behaviour*, 45, 243-264.

Valentino, N. A., Hutchings, V. L., Banks, A. J., & Davis, A. K. (2008). Is a worried citizen a good citizen? Emotions, political information seeking, and learning via the Internet. *Political Psychology*, 29(2), 247-273.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, 22(3), 213-224.

Vergeer, M. (2020). Political candidates' discussions on Twitter during election season: A network approach. In G. Bouvier & J. E. Rosenbaum (Eds.), *Twitter, the public sphere, and the chaos of online deliberation* (pp. 53-78). Palgrave.

von der Heyde, L., Haensch, A. C., & Wenz, A. (2024a). United in diversity? Contextual biases in LLM-based predictions of the 2024 European Parliament elections. *arXiv:2409.09045*.

von der Heyde, L., Haensch, A. C., & Wenz, A. (2024b). Vox Populi, Vox AI? Using language models to estimate German public opinion. *arXiv:2407.08563*.

Ward, S., & Lusoli, W. (2005). "From weird to wired": MPs, the Internet and representative politics in the UK. *Journal of Legislative Studies*, 11(1), 57-81.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Westphal, M. (2024). Affective polarization and political theory. In M. Torcal & E. Harteveld (Eds.), *Handbook of Affective Polarization*. Retrieved from `https://osf.io/86ds3/`

Widmann, T. (2022). Fear, hope, and COVID-19: Emotional elite rhetoric and its impact on the public during the first wave of the COVID-19 pandemic. *Political Psychology*, 43(5), 827-850.

Williams, N. W., Casas Salleras, A., Aslett, K., & Wilkerson, J. (2025). When conservatives see red but liberals feel blue: Labeler characteristics and the measurement of political concepts. *The Journal of Politics*.

Wlezien, C. (2024). News and public opinion: Which comes first? *Journal of Politics*, 86(1), 1-17.

Wlezien, C., & Soroka, S. (2024). Media reflect! Policy, the public, and the news. *American Political Science Review*, 118(3), 1563-1569.

Wojcik, S. (2017). Why legislative networks? Analyzing legislative network formation. *Political Science Research and Methods*, 7(3), 505-522.

Yam, K. C., Jackson, J. C., Barnes, C. M., Lau, J., Qin, X., & Lee, H. Y. (2020). The rise of COVID-19 cases is associated with support for world leaders. *Proceedings of the National Academy of Sciences*, 117(41), 25429-25433.

Yang, K. C., & Menczer, F. (2023). Accuracy and political bias of news source credibility ratings by large language models. *arXiv:2304.00228*.

Yoon, H. Y., & Park, W. H. (2014). Strategies affecting Twitter-based networking pattern of South Korean politicians: Social network analysis and exponential random graph model. *Quality & Quantity*, 48(1), 409-423.

Yoon, K., Pinkleton, B. E., & Ko, W. (2005). Effects of negative political advertising on voting intention: An exploration of the roles of involvement and source credibility in the development of voter cynicism. *Journal of Marketing Communications*, 11(2), 95-112.

Yuan, K., Qin, J., Wang, H., & Liu, Z. (2020). BERTweet: A pre-trained language model for English Tweets. *arXiv:2005.10200*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv:2305.03514*.

# A Appendix

## A.1 Survey Design and Descriptive Statistics

**Welcome:**

Welcome to this study! This is a study where you are asked to classify tweets on one dimension, for which you will receive extensive training and guidance. Before starting, this survey will ask more about your demographics, political preferences and general political interest. Although each of you has specific political leanings, it is crucial that you provide objective and unbiased scores. Thanks again for agreeing to participate in the study!

**Instructions and consent:**

Please take time to read the following information carefully and discuss it with others if you wish. Before taking part in this research project, it is important for you to understand why the research is being done and what it will involve.

This is a study examining attitudes and political views and has received ethical approval from the University of Oxford (ref: 1355365) It is generally interested in understanding how people make up their minds when forming political preferences. This study will ask you to classify tweets on one dimension, for which extensive training and guidance are offered. The principal investigator of this study is Giuliano Formisano, a PhD candidate in Social Data Science at the University of Oxford. Feel free to contact him at: giuliano.formisano@nuffield.ox.ac.uk if you have any questions, to provide feedback about your experience or if you are interested in the goals of this research. This information and consent form asks you to allow the researcher to record your responses and to use the data to further understand social and political values citizens might have.

Participation in this study is completely voluntary, it will only take a few minutes, and the data will be fully anonymous. It is up to you to decide whether you want to participate or not. If you decide not to participate, there will not be any negative consequences and no personal data will be stored. Please be aware that if you decide to participate, you may stop participating at any time, and you may decide not to answer any specific question. However, only those who will complete the full questionnaire will be reimbursed. The researcher will maintain the confidentiality of the research records or data, and all data will be stored at the University of Oxford.

If you have a concern about any aspect of this study, please contact the principal investigator. He will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, you may contact the Chair of the Research Ethics Committee at the University of Oxford at: drec@oii.ox.ac.uk, who will seek

to resolve the matter as soon as possible.

Responsible members of the research team, data repositories, and funders may be given access to data for monitoring and/or audit of the study to ensure the principal investigator complies with guidelines, or as otherwise required by law. The plan is to publish the results of this study in a scientific journal, and access to the anonymous data might be given to journal editors and reviewers.

Please note that you may only participate in this survey if you are 18 years of age or over.

If you have read the information above and agree to participate with the understanding that the data you submit will be processed accordingly, please check the relevant box below to get started.

Thank you in advance for your participation!

Here, respondents will have two boxes: Yes, I have read, understood, and I accept the rules stated above. I do not consent to participate in this survey.

Political Ideology and Demographic questions:

Generally speaking, do you think of yourself as being a Democrat, a Republican, or an Independent? [Strong Democrat, Democrat, Leaning Democrat, Independent, Leaning Republican, Republican, Strong Republican]

What is your gender? [Male, Female, Other]

What is your age? [Under 21 years old, 21–30 years old, 31–40 years old, 41–50 years old, 51–60 years old, Above 60 years old]

What is the highest level of education you have completed? [Some high school or less, High school diploma or GED, Some college, but no degree, Associates or technical degree, Bachelor's degree, Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.), Prefer not to say]

Generally speaking, to what extent would you say you are interested in politics and current affairs? [Very interested, Moderate interested, Slightly interested, Not interested at all]

**Coding instructions:**

Let's start the task! Remember, follow the instructions and classify the tweets accordingly. Response quality is important for this research.

Is the tweet written in a divisive manner? We give you three options: Divisive, Unifying and Neither Divisive nor Unifying.

DEFINITION Divisive: A tweet that expresses hostility toward the opposing side and favoritism toward one's own side.

Examples: @user over the last four years the Republicans have been destroying this country with policies that benefit the few, not the many. The Democratic agenda is the only one that can get us out of this mess"

The Democrats and their sham impeachment process with which @user you agree is the latest example of the un-American behavior of a party full of traitors, trying to remove the most honest president ever!"

DEFINITION Unifying: A tweet that emphasizes the need for collaboration with the other side.

Examples: "We've had our differences with President Trump but all this should be put aside now. We need to come together as a nation and build our future"

DEFINITION Neither Divisive nor Unifying: A tweet in which none of the above apply.

Let's see some examples in practice

Figure 1: Coding Instructions - Example 1

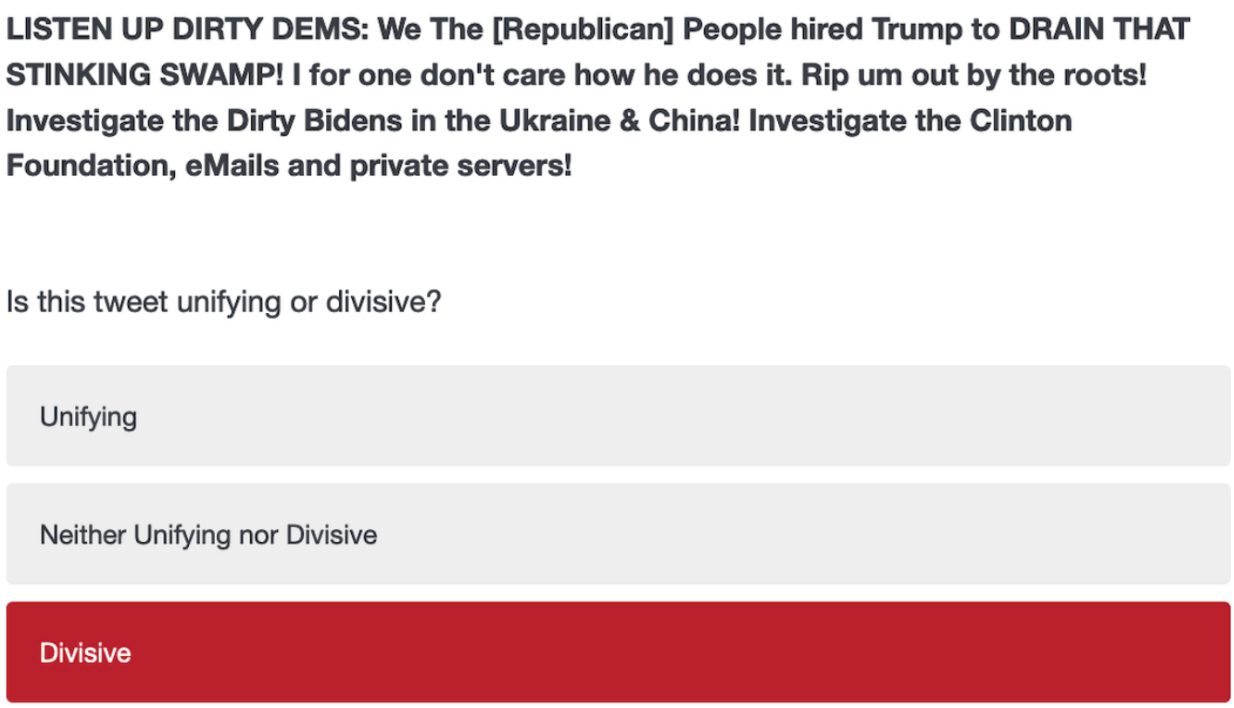


Figure 2: Coding Instructions - Example 2




Figure 3: Coding Instructions - Example 3

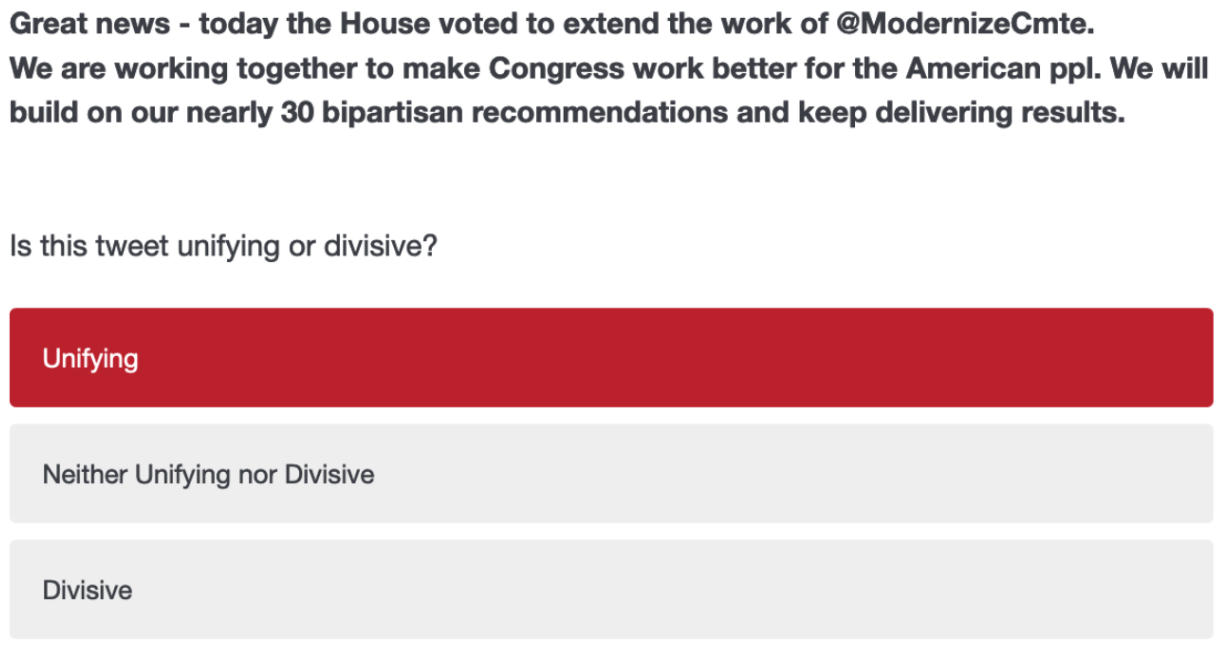
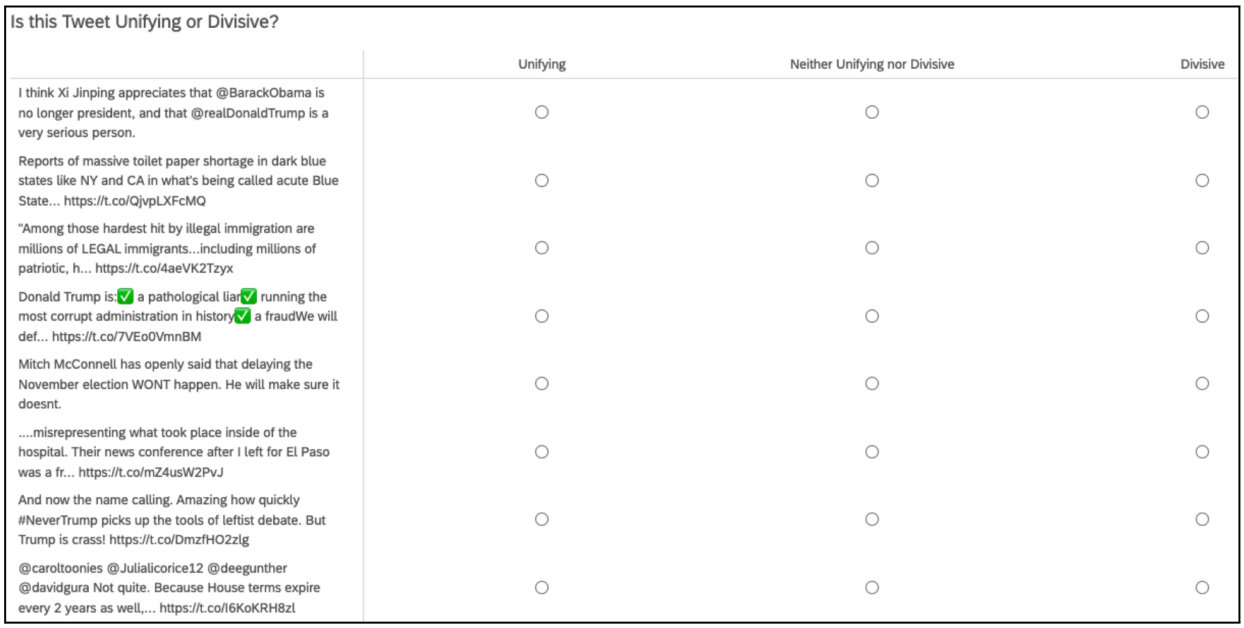


Figure 4: Example of a coding task


Is this Tweet Unifying or Divisive?

| | Unifying | Neither Unifying nor Divisive | Divisive |
|---|---|---|---|
| I think Xi Jinping appreciates that @BarackObama is no longer president, and that @realDonaldTrump is a very serious person. | ○ | ○ | ○ |
| Reports of massive toilet paper shortage in dark blue states like NY and CA in what's being called acute Blue State... https://t.co/QjvpLXFcMQ | ○ | ○ | ○ |
| "Among those hardest hit by illegal immigration are millions of LEGAL immigrants...including millions of patriotic, h... https://t.co/4aeVK2Tzyx | ○ | ○ | ○ |
| Donald Trump is:✅ a pathological liar✅ running the most corrupt administration in history✅ a fraudWe will def... https://t.co/7VEo0VmnBM | ○ | ○ | ○ |
| Mitch McConnell has openly said that delaying the November election WONT happen. He will make sure it doesnt. | ○ | ○ | ○ |
| ....misrepresenting what took place inside of the hospital. Their news conference after I left for El Paso was a fr... https://t.co/mZ4usW2PvJ | ○ | ○ | ○ |
| And now the name calling. Amazing how quickly #NeverTrump picks up the tools of leftist debate. But Trump is crass! https://t.co/DmzfHO2zlg | ○ | ○ | ○ |
| @caroltoonies @Julialicorice12 @deegunther @davidgura Not quite. Because House terms expire every 2 years as well,... https://t.co/I6KoKRH8zl | ○ | ○ | ○ |

Table 5: Summary Statistics of Coders

| Coders | | N Democrat | N Republican | N Independents | N |
|---|---|---|---|---|---|
| **Partisanship** | | 1,291 | 1,190 | 1,108 | 3,589 |
| **Gender** | **Male** | 563 | 538 | 550 | 1,651 |
| | **Female** | 705 | 647 | 551 | 1,903 |
| | **Other** | 23 | 5 | 7 | 35 |
| **Age** | **Under 21 years old** | 23 | 33 | 34 | 90 |
| | **21-30 years old** | 287 | 256 | 244 | 787 |
| | **31-40 years old** | 284 | 302 | 317 | 903 |
| | **41-50 years old** | 219 | 198 | 246 | 663 |
| | **51-60 years old** | 177 | 148 | 148 | 473 |
| | **Above 60 years old** | 97 | 96 | 106 | 299 |
| **Education** | **Associates or technical degree** | 99 | 109 | 122 | 330 |
| | **Bachelor's degree** | 415 | 399 | 403 | 1,217 |
| | **Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)** | 299 | 248 | 211 | 758 |
| | **High school diploma or GED** | 99 | 105 | 139 | 343 |
| | **Some college, but no degree** | 144 | 191 | 213 | 548 |
| | **Some high school or less** | 6 | 6 | 7 | 19 |
| **Interest in Politics** | **Moderate interested** | 517 | 441 | 514 | 1,472 |
| | **Not interested at all** | 24 | 31 | 45 | 100 |
| | **Slightly interested** | 196 | 180 | 232 | 608 |
| | **Very interested** | 325 | 405 | 302 | 1,032 |

## A.2 Dis/Agreements Across Coders

Figure 5: Examples of tweets with disagreements across coders

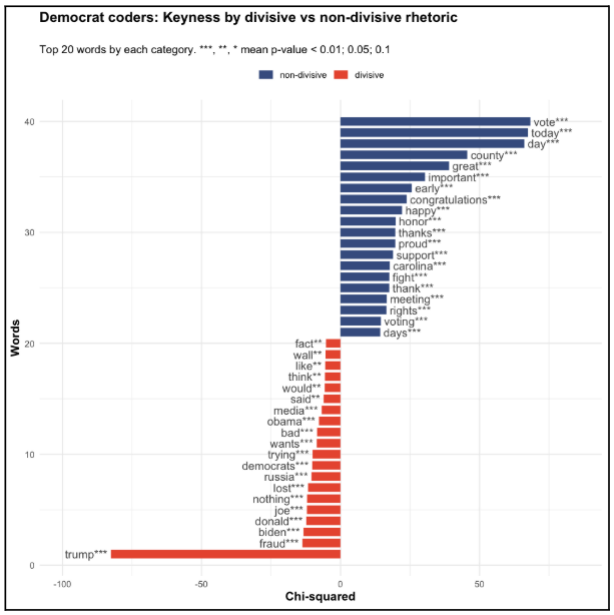
| Tweet | Poster | Coders | | | |
|---|---|---|---|---|---|
| | | Democrat | Republican | Combined | GPT4 |
| Mitch McConnell has openly said that delaying the November election WONT happen. He will make sure it doesnt. | Unknown | 0 | 1 | 0 | 0 |
| ....misrepresenting what took place inside of the hospital. Their news conference after I left for El Paso was a fr… https://t.co/mZ4usW2PvJ | Republican | 1 | 1 | 1 | 0 |
| @LpoolLefty11 it's a constitutional law for states like michigan and pennsylvania and wisconsin that ballots cannot… https://t.co/qiXIbDfl4E | Unknown | 0 | 1 | 0 | 0 |
| Andriy Derkach is the Ukrainian pushing Biden/Ukraine propaganda who "magically" came into possession of tapes of B… https://t.co/lgCOwspnwv | Democrat | 1 | 1 | 1 | 0 |
| Cruz fired a staffer for a Facebook post.Trump campaign manager accused of assault, and he blames the victim. | Democrat | 1 | 1 | 1 | 0 |
| ICYMI: My Q&A with @realDonaldTrump's nominee, John Ryder of #Memphis, to be a member of the Board of Directors of… https://t.co/9865PKhh35 | Republican | 0 | 1 | 1 | 0 |
| This kind of behavior should land you in an ethics investigation, not earn you a spot in the Cabinet: https://t.co/U4V6iZlAHj | Democrat | 1 | 1 | 1 | 0 |
| Supreme Court declines to change Florida voting rule aimed at easing convicted felons' ability to vote in 2020 elec… https://t.co/ssClMq6xCw | Unknown | 1 | 1 | 1 | 0 |
| .@RealDonaldTrump told the vice president not to reach out to #DemGovs who aren't "appreciative" of the administrat… https://t.co/K85EDZJN8t | Democrat | 1 | 1 | 1 | 0 |
| The media dishonesty is so over the top. We actually need good journalism right now-asking honest questions and obj… https://t.co/qCDbgybFTY | Republican | 1 | 1 | 1 | 0 |
| Your last chance to laugh, before we all cry over the election.... https://t.co/Rr4dZEU9wQ | Unknown | 1 | 0 | 1 | 0 |
| @Jiillyy @JordanChariton It's interesting with Bernie fans who@lost last election to Clinton cried foul on rules th… https://t.co/E3Hk8PxtJ9 | Unknown | 1 | 1 | 1 | 0 |
| @TrumpsMeltDown @realDonaldTrump And how long have you been saying that? ROFL😆🤣😂 | Republican | 1 | 1 | 1 | 0 |
| Only 50 days until Election Day! Can you chip in $50 to help #TeamBurgess ensure an Election Day win? https://t.co/lHu3H1v7Wc | Republican | 1 | 0 | 0 | 0 |
| I would only say if you think he'll do more harm than Hillary Clinton, you should look again, really close. https://t.co/390mlS5JQp | Republican | 1 | 1 | 1 | 0 |
| @marthamaccallum Since the election, the Media has continually said that Trump is divisive. Now we know that the Me… https://t.co/NwKQaSfZ1L | Unknown | 1 | 1 | 1 | 0 |
| In #MEpolitics: 56% of voters oppose Trump's use of feds to suppress protests. 56% want Congress to intervene, 58%… https://t.co/CPjtuJjL6S | Democrat | 1 | 1 | 1 | 0 |
| Trump's affinity for strongmen is strengthened every time he looks in the mirror. Combined with his explicit underm… https://t.co/Sm6VECSVlv | Democrat | 1 | 0 | 1 | 1 |
| I know @JimDeMint and his team have been invaluable to President Trump in the transition to actually running the government. (3/3) | Republican | 1 | 0 | 1 | 0 |
| #SC04 & @IndivisibleSC04 & @Indivisible_SC4! We're looking for voters in your district to send us a video of why yo… https://t.co/DfxepFX2dT | Democrat | 1 | 0 | 0 | 0 |

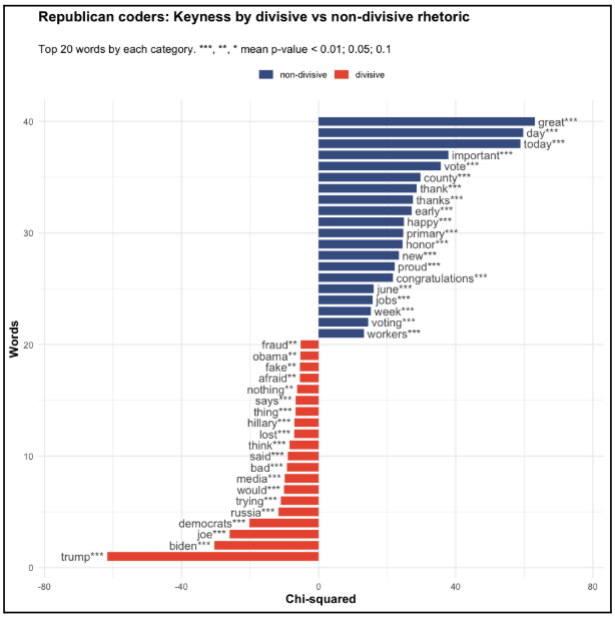
Figure 6: Examples of tweets with agreements across coders

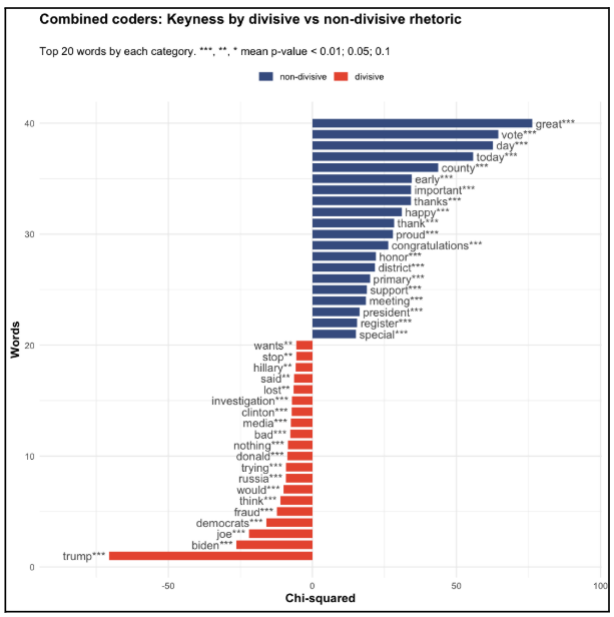
| Tweet | Poster | Coders | | | |
|---|---|---|---|---|---|
| | | Democrat | Republican | Combined | GPT4 |
| I think Xi Jinping appreciates that @BarackObama is no longer president, and that @realDonaldTrump is a very serious person. | Republican | 1 | 1 | 1 | 1 |
| Reports of massive toilet paper shortage in dark blue states like NY and CA in what's being called acute Blue State… https://t.co/QjvpLXFcMQ | Republican | 1 | 1 | 1 | 1 |
| "Among those hardest hit by illegal immigration are millions of LEGAL immigrants…including millions of patriotic, h… https://t.co/4aeVK2Tzyx | Republican | 0 | 0 | 0 | 0 |
| Donald Trump is:✅ a pathological liar✅ running the most corrupt administration in history✅ a fraudWe will def… https://t.co/7VEo0VmnBM | Democrat | 1 | 1 | 1 | 1 |
| And now the name calling. Amazing how quickly #NeverTrump picks up the tools of leftist debate. But Trump is crass! https://t.co/DmzfHO2zlg | Republican | 1 | 1 | 1 | 1 |
| @caroltoonies @Julialicorice12 @deegunther @davidgura Not quite. Because House terms expire every 2 years as well,… https://t.co/l6KoKRH8zl | Unknown | 0 | 0 | 0 | 0 |
| Here we go again: the Trump administration is peddling anti-Blackness and Islamophobia to divide and distract us. B… https://t.co/K56P8PDI8p | Democrat | 1 | 1 | 1 | 1 |
| Schumer doesn't want FBI to investigate biden laptop before the election! We all know why, it's a death sentence fo… https://t.co/YYKJpyU3p1 | Unknown | 1 | 1 | 1 | 1 |
| @RepMattGaetz He's a loser. He also just LOST the biggest election of his life.Trump=LOSER(and these are FACTS) | Unknown | 1 | 1 | 1 | 1 |
| I'm hopeful.But here's the problem:Kim won. He got legitimacy. He got respect. He got time. He bought himself l… https://t.co/agqp9ckD3a | Republican | 0 | 0 | 0 | 0 |
| Awesome! And this is a district Trump won in 2016. Congratulations, @jimhaadsma, @michigandems, and @mihousedems! https://t.co/d2VNlBJs5Y | Democrat | 0 | 0 | 0 | 0 |
| @hollysoriano @JennieSTaer No! the election is over, "fraud" was a joke 250k Americans deas from a virus and we hav… https://t.co/NL0WvnmpXv | Unknown | 1 | 1 | 1 | 1 |
| Cutting off funding to the WHO at the peak of the pandemic, just to score cheap political points, will go down as o… https://t.co/htfrxDDy38 | Democrat | 1 | 1 | 1 | 1 |
| This week, @SenateDems and some Senate Republicans stood together with the Constitution, the law, and the people we… https://t.co/xAk8x1vWFC | Democrat | 0 | 0 | 0 | 0 |
| @Twitter Why can't you block @realDonaldTrump Twitter account since he won't stop posting lies about the election a… https://t.co/Fv4hGRqb9W | Unknown | 1 | 1 | 1 | 1 |
| Pete Sessions and Donald Trump are sabotaging our healthcare system and raising costs for you and your family. #TX32 https://t.co/qd8LtvUTld | Democrat | 1 | 1 | 1 | 1 |
| The failing @NYDailyNews, destroyed by little Morty Zuckerman, is preparing to close and save face by going online. It's dead! | Republican | 1 | 1 | 1 | 1 |
| RACE UPDATE: Ohio is officially a TOSS UP for 2020!The race for Ohio is heating up. Both @JoeBiden and President… https://t.co/qVFHV87oC6 | Democrat | 0 | 0 | 0 | 0 |
| #100days = "Trump failed to be the champion for America's workers that he promised he would be" -- @SenSchumer https://t.co/CQ9PsaMDJ0 | Democrat | 1 | 1 | 1 | 1 |
| @RepDougCollins @LouDobbs Need to be in Georgia state house fixing voter fraud. Can't have fair election without it. | Unknown | 0 | 0 | 0 | 0 |

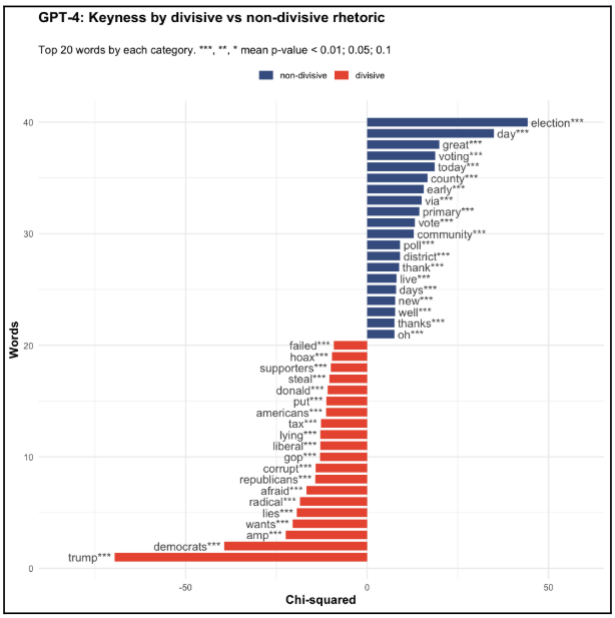
## A.3 Keyness Analysis by Coders' types

Figure 7: Keyness Analysis by Coders for Non- vs Divisive rhetoric



## A.4 Comparison against Simchon et al. (2022)

The comparison of Simchon et al.'s polarisation predictions with those of human coders and GPT-4o has been conducted using t-tests, point-biserial correlation, Receiver Operating Characteristic (ROC) curve analysis, and the determination of the optimal binary classification threshold. The results are as follows:

Simchon et al. vs Combined Coders: There is a highly significant difference between the polarisation scores (t-statistic = 9.47, p-value < 0.0001). The point-biserial correlation is moderate (r = 0.1705, p-value < 0.0001). The AUC for the ROC curve is relatively strong (0.603), with an optimal binary classification threshold of 1.0.

Simchon et al. vs Democrat Coders: The t-test indicates a highly significant difference (t-statistic = 7.98, p-value < 0.0001). The correlation is moderate (r = 0.1443, p-value < 0.0001). The AUC is slightly lower than the Combined Coders comparison (0.594), with the best threshold for classification at 1.0.

Simchon et al. vs Republican Coders: A highly significant difference is observed between the groups (t-statistic = 9.89, p-value < 0.0001). The correlation is moderate (r = 0.1778, p-value < 0.0001). The AUC is comparable to the previous comparisons (0.601), and the best threshold for classification is 1.0.

Simchon et al. vs GPT-4o: The most significant difference is observed in this comparison (t-statistic = 14.24, p-value < 0.0001). The point-biserial correlation is the highest among comparisons (r = 0.2517, p-value < 0.0001), and the AUC is also the highest (0.624). The best threshold for binary classification is 1.0.

In summary, Simchon et al.'s polarisation predictions show highly significant differences with all comparator groups, including Combined Coders, Democrat Coders, Republican Coders, and GPT-4o. Unlike the previous results, the Republican Coders now show a significant difference. The best classification threshold is consistently 1.0 across all comparisons, with GPT-4o showing the strongest performance based on the AUC. Below, I display the full results.

Figure 8: Distribution of Simchon et al. vs Combined Coders' Polarisation Predictions

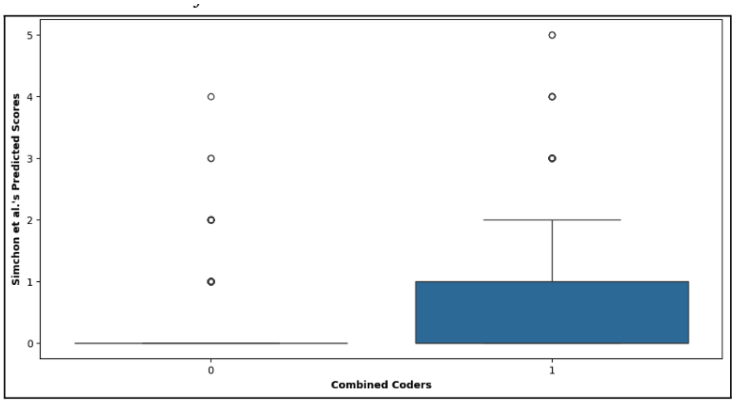
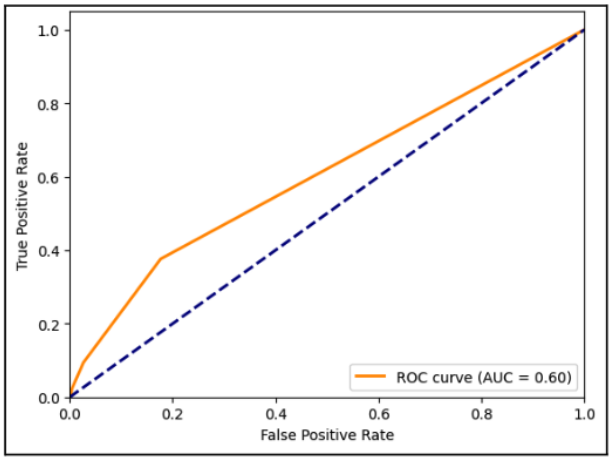



Figure 9: Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs Combined Coders




Simchon et al. vs Combined Coders

- T-test result: t-statistic = 9.472460144312922, p-value = 5.35718666519025e-21
- Point-biserial correlation: r = 0.17046817185506186, p-value = 5.357186665164164e-21
- AUC: 0.6031054741906356
- Best threshold for matching binary classification: 1.0

Figure 10: Distribution of Simchon et al. vs Democrat Coders' Polarisation Predictions

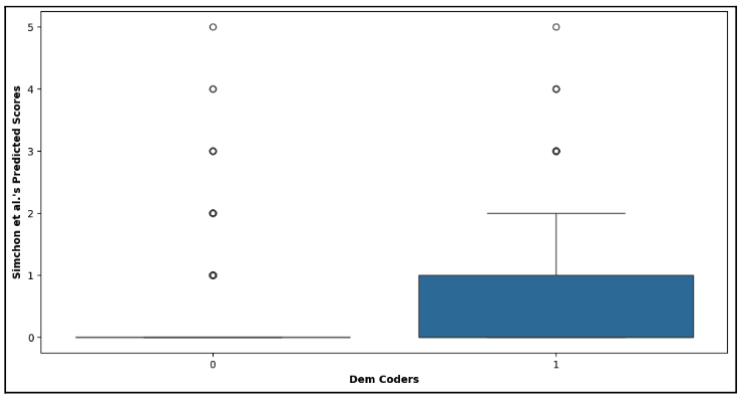
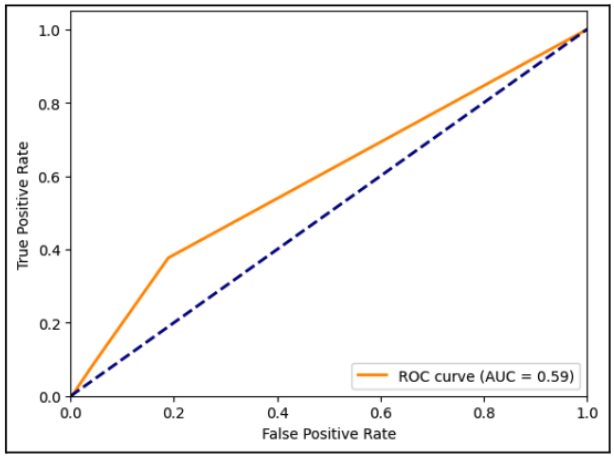



Figure 11: Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs Democrat Coders




Simchon et al. vs Dem Coders

- T-test result: t-statistic = 7.984923348373899, p-value = 1.9848196523300867e-15
- Point-biserial correlation: r = 0.14430628613801005, p-value = 1.9848196523195187e-15
- AUC: 0.5936409993361897
- Best threshold for matching binary classification: 1.0

Figure 12: Distribution of Simchon et al. vs Rep Coders' Polarisation Predictions

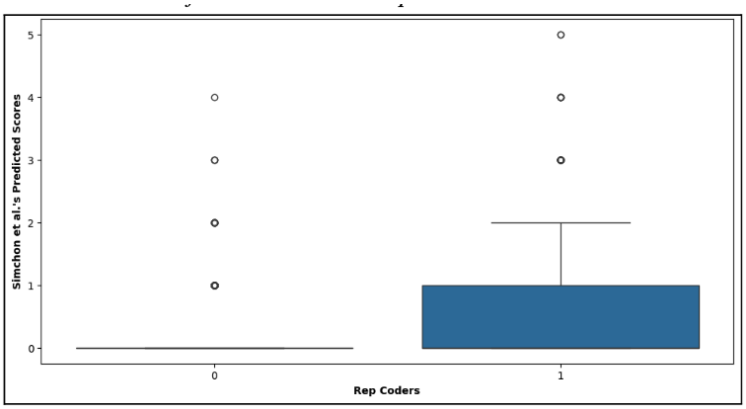
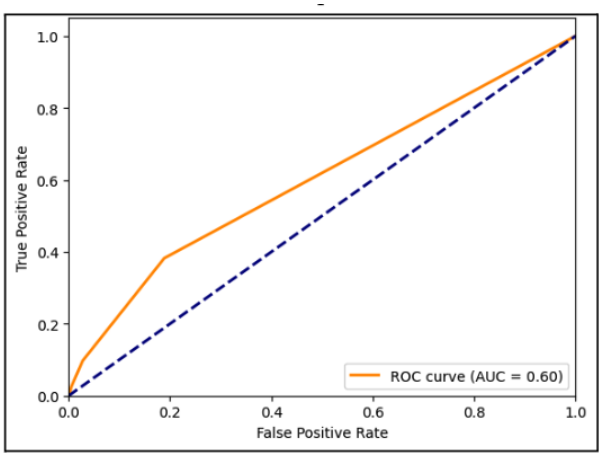



Figure 13: Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs Republican Coders




Simchon et al. vs Rep Coders

- T-test result: t-statistic = 9.891348001183811, p-value = 1.007353840084555e-22
- Point-biserial correlation: r = 0.17777321008023322, p-value = 1.0073538400791341e-22
- AUC: 0.6007647000099832
- Best threshold for matching binary classification: 1.0

Figure 14: Distribution of Simchon et al. vs GPT-4o's Polarisation Predictions

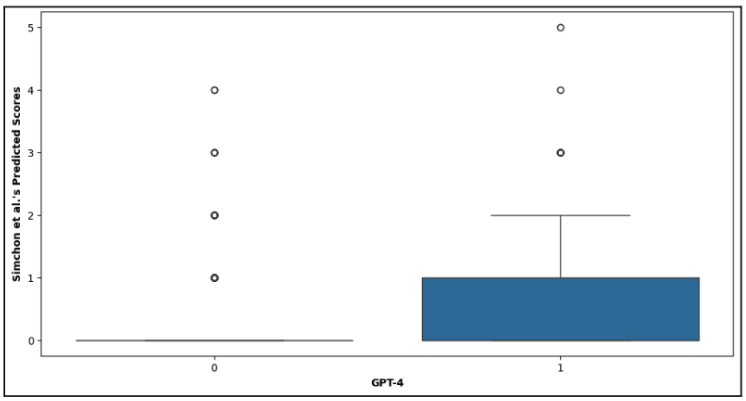
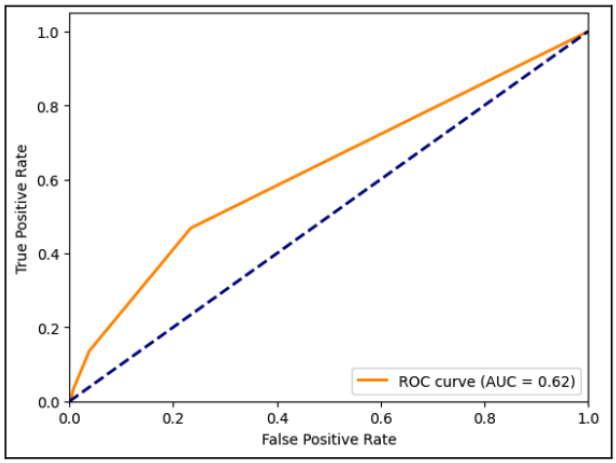



Figure 15: Receiver Operating Characteristic (ROC) Curve - Simchon et al. vs GPT-4o




Simchon et al. vs GPT-4o:

- T-test result: t-statistic = 14.239585599096792, p-value = 1.4306104744920922e-44
- Point-biserial correlation: r = 0.2516925737495741, p-value = 1.4306104744846869e-44
- AUC: 0.6242182173501145
- Best threshold for matching binary classification: 1.0

## A.5 Topic Modelling (Paper 1)

Using BERTopic (Grootendorst, 2022), which is discussed in detail in Paper 2, I study the topics mentioned in the tweets used to train the classifiers. I also provide the descriptive statistics of the predicted topics. Given the limited number of topics discussed, topic reduction has not been performed.

Descriptive statistics

- count: 1,633
- mean: 0.78
- std: 0.24

Table 6: Summary of the Topics discussed

| Topic | Renamed Topic | Count |
|---|---|---|
| 0_biden_obama_trump_president | US Presidential Politics | 772 |
| 1_ballot_vote_ballots_voting | Voting & Elections | 215 |
| 2_election_voter_fraud_fraudulent | Election Fraud & Integrity | 205 |
| 3_trumpcare_obamacare_healthcare_medicare | Healthcare & Insurance | 136 |
| 4_tariffs_taxes_tax_economy | Taxes & Economy | 89 |
| 5_refugees_immigration_asylum_children | Immigration & Refugees | 46 |
| 6_justices_scalia_nominee_judges | Supreme Court & Judicial Appointments | 45 |
| 7_abortion_parenthood_rights_nra | Abortion & Reproductive Rights | 32 |
| 8_racists_racist_racismmov_racism | Racism & Racial Issues | 28 |
| 9_postal_usps_mail_emails | Postal Service & Mail-in Voting | 17 |
| 10_israeli_israel_palestinians_gaza | Israel & Palestine Conflict | 13 |
| 11_tuition_college_colleges_student | Higher Education & Student Loans | 12 |
| 12_democrats_socialist_democrat_dems | Democrats vs Republicans | 12 |
| 13_maxinewaters_trump_pac_maxine | Maxine Waters & Trump Feud | 11 |

## A.6 Summary Tables of Classifiers' Performances

Figure 16: Democrat coders - Summary table for "Divisiveness"

| Info | Model | Seed | Train/Val/Test Split | Accuracy | Weighted F-score | AUC |
|---|---|---|---|---|---|---|
| Coding: 0 = non-divisive; 1 = divisive<br><br>Distribution: 0 = 730 tweets; 1 = 2,270 tweets | BERTweet | 676 | 60% training 20% validation 20% testing | 85% | 86% | 0.81 |
| | | 291 | | 85% | 86% | 0.82 |
| | | 887 | | 87% | 87% | 0.81 |
| | | 188 | 70% training 15% validation 15% testing | 86% | 86% | 0.81 |
| | | 206 | | 85% | 85% | 0.83 |
| | | 823 | | 84% | 84% | 0.82 |
| | | 52 | 80% training 10% validation 10% testing | 84% | 85% | 0.79 |
| | | 726 | | 87% | 87% | 0.79 |
| | | 721 | | 86% | 86% | 0.80 |
| | roBERTa | 580 | 60% training 20% validation 20% testing | 87% | 87% | 0.82 |
| | | 785 | | 85% | 86% | 0.83 |
| | | 903 | | 86% | 86% | 0.82 |
| | | 307 | 70% training 15% validation 15% testing | 82% | 83% | 0.83 |
| | | 115 | | 86% | 86% | 0.82 |
| | | 550 | | 86% | 87% | 0.82 |
| | | 824 | 80% training 10% validation 10% testing | 86% | 86% | 0.79 |
| | | 524 | | 86% | 87% | 0.83 |
| | | 833 | | 88% | 88% | 0.81 |
| | Frozen roBERTa | 504 | 80% training 10% validation 10% testing | 74% | 63% | 0.5 |
| | | 336 | | 75% | 64% | 0.5 |
| | | 940 | | 77% | 67% | 0.5 |
| | Majority Vote Baseline | 606 | 80% training 20% testing | 77% | 67% | 0.5 |
| | | 19 | | 74% | 62% | 0.5 |
| | | 171 | | 77% | 67% | 0.5 |

Figure 17: Republican coders - Summary table for “Divisiveness”

| Info | Model | Seed | Train/Val/Test Split | Accuracy | Weighted F-score | AUC |
|---|---|---|---|---|---|---|
| Coding:<br>0 = non-divisive;<br>1 = divisive<br><br>Distribution:<br>0 = 795 tweets;<br>1 = 2,205 tweets | BERTweet | 285 | 60% training<br>20% validation<br>20% testing | 85% | 85% | 0.84 |
| | | 264 | | 88% | 88% | 0.83 |
| | | 832 | | 86% | 86% | 0.79 |
| | | 804 | 70% training<br>15% validation<br>15% testing | 84% | 84% | 0.78 |
| | | 120 | | 86% | 85% | 0.78 |
| | | 616 | | 84% | 85% | 0.82 |
| | | 230 | 80% training<br>10% validation<br>10% testing | 88% | 88% | 0.87 |
| | | 403 | | 87% | 86% | 0.77 |
| | | 543 | | 80% | 81% | 0.81 |
| | roBERTa | 366 | 60% training<br>20% validation<br>20% testing | 86% | 87% | 0.83 |
| | | 766 | | 83% | 83% | 0.75 |
| | | 151 | | 82% | 82% | 0.78 |
| | | 668 | 70% training<br>15% validation<br>15% testing | 85% | 84% | 0.80 |
| | | 919 | | 84% | 85% | 0.81 |
| | | 692 | | 86% | 87% | 0.84 |
| | | 467 | 80% training<br>10% validation<br>10% testing | 85% | 86% | 0.84 |
| | | 600 | | 87% | 87% | 0.83 |
| | | 155 | | 84% | 84% | 0.83 |
| | Frozen BERTweet | 594 | 80% training<br>10% validation<br>10% testing | 77% | 69% | 0.56 |
| | | 415 | | 78% | 74% | 0.63 |
| | | 439 | | 84% | 81% | 0.66 |
| | Majority Vote Baseline | 550 | 80% training<br>20% testing | 72% | 61% | 0.5 |
| | | 52 | | 74% | 62% | 0.5 |
| | | 984 | | 76% | 66% | 0.5 |

Figure 18: Combined coders - Summary table for “Divisiveness”

| Info | Model | Seed | Train/Val/Test Split | Accuracy | Weighted F-score | AUC |
|---|---|---|---|---|---|---|
| Coding:<br>0 = non-divisive;<br>1 = divisive<br><br>Distribution:<br>0 = 679 tweets;<br>1 = 2,321 tweets | BERTweet | 968 | 60% training<br>20% validation<br>20% testing | 84% | 85% | 0.82 |
| | | 854 | | 89% | 89% | 0.82 |
| | | 929 | | 88% | 87% | 0.77 |
| | | 544 | 70% training<br>15% validation<br>15% testing | 90% | 89% | 0.82 |
| | | 623 | | 86% | 86% | 0.84 |
| | | 301 | | 89% | 89% | 0.83 |
| | | 204 | 80% training<br>10% validation<br>10% testing | 91% | 90% | 0.85 |
| | | 704 | | 88% | 88% | 0.83 |
| | | 116 | | 90% | 90% | 0.87 |
| | roBERTa | 293 | 60% training<br>20% validation<br>20% testing | 89% | 88% | 0.82 |
| | | 384 | | 86% | 85% | 0.77 |
| | | 429 | | 87% | 87% | 86% |
| | | 331 | 70% training<br>15% validation<br>15% testing | 89% | 89% | 0.85 |
| | | 566 | | 87% | 87% | 0.84 |
| | | 530 | | 86% | 87% | 0.84 |
| | | 199 | 80% training<br>10% validation<br>10% testing | 85% | 85% | 0.82 |
| | | 552 | | 91% | 91% | 0.84 |
| | | 656 | | 89% | 87% | 0.72 |
| | Frozen roBERTa | 578 | 80% training<br>10% validation<br>10% testing | 77% | 67% | 0.50 |
| | | 76 | | 79% | 70% | 0.50 |
| | | 250 | | 78% | 68% | 0.50 |
| | Majority Vote Baseline | 885 | 80% training<br>20% testing | 79% | 70% | 0.5 |
| | | 780 | | 80% | 71% | 0.5 |
| | | 254 | | 79% | 70% | 0.5 |

Figure 19: GPT-4o Coded - Summary table for "Divisiveness"

| Info | Model | Seed | Train/Val/Test Split | Accuracy | Weighted F-score | AUC |
|---|---|---|---|---|---|---|
| Coding:<br>0 = non-divisive;<br>1 = divisive<br><br>Distribution:<br>0 = 1,757 tweets;<br>1 = 1,243 tweets | BERTweet | 575 | 60% training<br>20% validation<br>20% testing | 82% | 82% | 0.82 |
| | | 76 | | 82% | 82% | 0.82 |
| | | 667 | | 82% | 82% | 0.82 |
| | | 912 | 70% training<br>15% validation<br>15% testing | 86% | 86% | 0.86 |
| | | 149 | | 84% | 84% | 0.83 |
| | | 333 | | 83% | 83% | 0.81 |
| | | 883 | 80% training<br>10% validation<br>10% testing | 81% | 81% | 0.80 |
| | | 951 | | 83% | 83% | 0.83 |
| | | 621 | | 80% | 81% | 0.82 |
| | roBERTa | 628 | 60% training<br>20% validation<br>20% testing | 81% | 81% | 0.80 |
| | | 141 | | 82% | 82% | 0.82 |
| | | 546 | | 80% | 80% | 0.82 |
| | | 955 | 70% training<br>15% validation<br>15% testing | 85% | 85% | 0.85 |
| | | 724 | | 85% | 85% | 0.84 |
| | | 366 | | 81% | 81% | 0.79 |
| | | 889 | 80% training<br>10% validation<br>10% testing | 83% | 83% | 0.83 |
| | | 609 | | 78% | 78% | 0.79 |
| | | 286 | | 83% | 83% | 0.83 |
| | Frozen BERTweet | 400 | 70% training<br>15% validation<br>15% testing | 68% | 68% | 0.66 |
| | | 4 | | 72% | 71% | 0.69 |
| | | 545 | | 73% | 73% | 0.72 |
| | Majority Vote Baseline | 320 | 80% training<br>20% testing | 59% | 44% | 0.50 |
| | | 142 | | 58% | 43% | 0.50 |
| | | 704 | | 59% | 44% | 0.50 |

GPT-4o version: gpt-4o-2024-11-20, released on November 20, 2024

## A.7 Ego vs Socio-tropic considerations

I now examine two more sets of arguments: ego vs socio-tropic considerations. For the former, I build on previous political economy research outside political communication, in particular following work by Margalit (2013), and more recently by Schaffer and Spilker (2019). The authors explore how ego-tropic considerations may outweigh individuals' prior ideological dispositions, resulting in a significant, even if temporary, shift in political preferences. Schaffer and Spilker (2019) study the relationship between ego versus socio-tropic considerations to understand individuals' trade preferences. Their survey shows that there is a net difference in these two considerations, with self-interest being more important than socio-interest in predicting preferences. Notably, the authors find little evidence for a socio-tropic view of trade openness given that individuals may not react to country-level information in isolation, which contrasts their prior egocentric views.

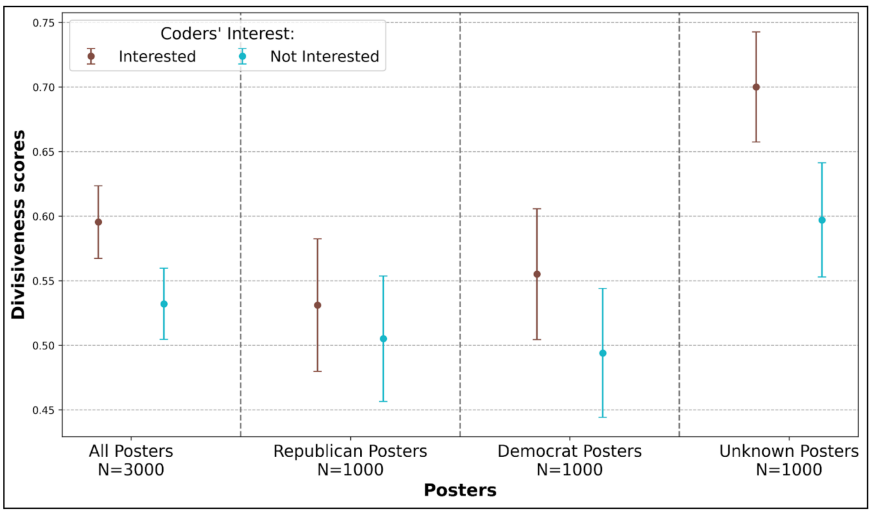
On the socio-tropic aspect, Mansfield and Mutz (2009) dismiss self-interest as the exclusive factor shaping preferences. They contend that individuals form their views, relying on collective information at the national level. In the context of political economy, Mansfield and Mutz (2009) find that people decide their stance by considering broader informational cues, such as the overall benefits or drawbacks the nation might experience from engaging in international trade, rather than solely evaluating trade's impact on their personal interests. Therefore, individuals utilise the available data regarding trade's effects on the national scale to supplement their limited understanding of its personal implications (ibidem). Returning to political communication, I would expect individuals who are very interested in politics to be more likely to know what users from another party may interpret a text as polarising.

Another question that I am interested in investigating is whether coders perceive polarisation based on their own perspective (ego-tropic) or based on what they believe could be perceived as polarising by users from other parties (socio-tropic). To study such cases, I utilise users' interest in politics as a proxy, which I have received on a 4-point scale ranging from "not interested" to "very interested" in politics. One would expect a person with high interest in politics to be more likely to know what a user from another party would consider as polarising (socio-tropic). Whereas, an individual less interested in politics may not know how an individual from the opposite side of the political spectrum would perceive as polarising, thus basing their labelling on their own perception (ego-tropic). In other words, those who are more interested in politics are more likely to know what users' from another party may interpret text as polarising.

Figure 20 compares coders' judgments of divisiveness based on their level of political interest (used as a proxy for socio-tropic and ego-tropic considerations) across different groups of tweet posters (all, Republican, Democrat, and unknown). In this analysis, coders who are interested in politics are assumed to make socio-tropic evaluations, while those not interested are assumed to

make ego-tropic evaluations. The results show substantial differences in polarisation judgments between politically interested and uninterested coders when analysing tweets from all posters and from unknown posters. However, these differences are not significant when focusing specifically on tweets posted by Democrats or Republicans.

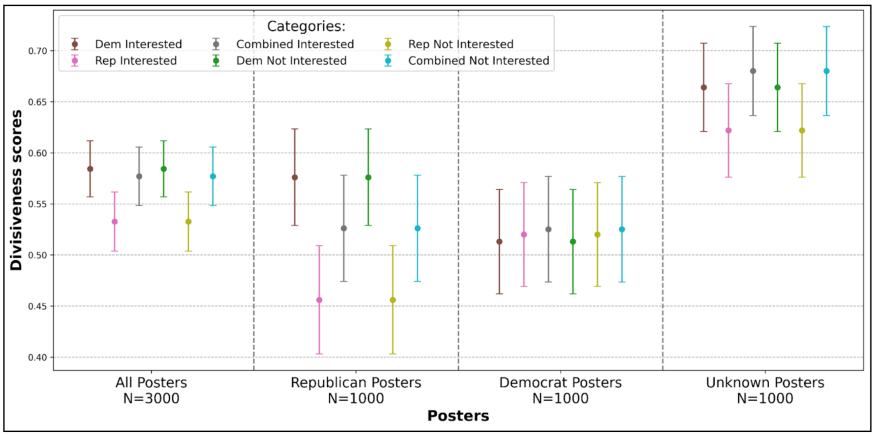
Figure 20: Comparison of Divisiveness Scores across Ego (Not-interested) vs Socio-tropic (Interested) Considerations




*Note: Coders' Political Interest and Tweets' Posters considered. Mean and 95% CI provided.*

Figure 21 deepens this analysis by accounting for both coders' political interest and partisanship. When isolating these factors, differences between high- and low-interest groups fade, as confidence intervals overlap substantially. However, combining interest and partisanship uncovers a key nuance for tweets posted by Republicans: partisanship dominates. For instance, a Republican tweet is rated as highly divisive by interested Democrats but minimally divisive by interested Republicans. This partisan split persists across interest levels, underscoring its primacy. On the contrary, for tweets posted by Democrats, unknown users, or when pooling all tweets together, neither interest nor partisanship yields significant perceptual divides.

Figure 21: Comparison of Divisiveness Scores across Ego (Not-interested) vs Socio-tropic (Interested) Considerations by Partisanship

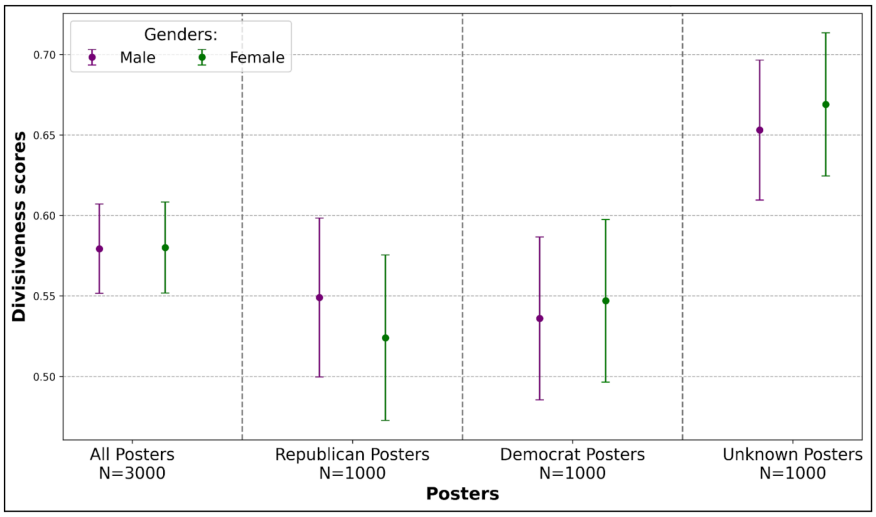



*Note: Coders' Political Interest and Partisanship, and Tweets' Posters considered. Mean and 95% CI provided.*

## A.8 Gender effect on Perception

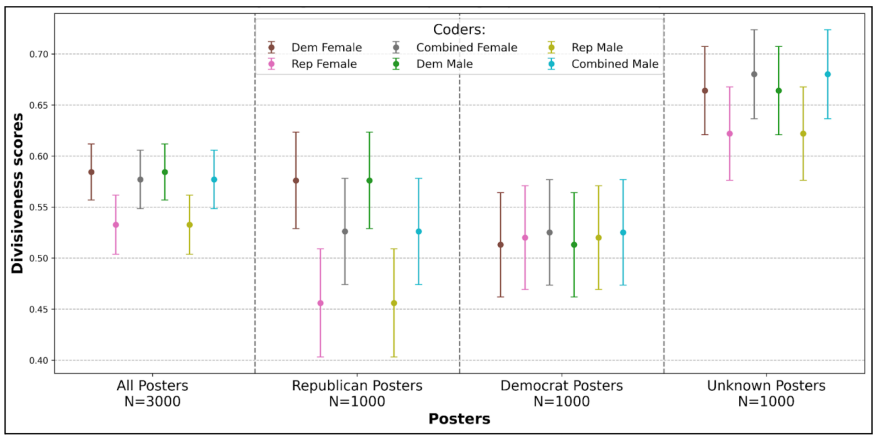
Building on Williams et al. (2025), I explore whether gender shapes how people perceive divisiveness in tweets. The analysis below points to partisanship—not gender—as the main driver of these perceptions. For example, when comparing how men and women rate the divisiveness of a tweet about healthcare reform, I find no significant difference between their responses. Replicating the distributional analyses as described above, this pattern holds across different posters. However, when I consider both gender and partisanship, the results shift for tweets posted by Republicans. In those cases, both male and female Democratic coders rate it as highly divisive, while both male and female Republican coders see it as less divisive. Here, gender does not matter—there are no significant differences between male and female Democrats or Republicans. Instead, it is partisanship that shapes the perception of divisiveness within each gender. Notably, this partisan effect disappears when looking at tweets from Democrats, tweets from unknown posters, or when pooling all tweets together. In those cases, neither gender nor partisanship leads to significant differences in how divisiveness is perceived.

Figure 22: Comparison of Divisiveness Scores by Gender




*Note: Coders' Gender and Tweets' Posters considered. Mean and 95% CI provided.*

Figure 23: Comparison of Divisiveness Scores by Gender and Partisanship




*Note: Coders' Gender and Partisanship, and Tweets' Posters considered. Mean and 95% CI provided.*

## A.9 Key Dates for Time-Series Annotations

- 2015

  June 26: Supreme Court legalizes same-sex marriage nationwide

  October 1: Oregon College Shooting & Gun Control Debate

  December 2: San Bernardino Shooting

- 2016

  January 2: Occupation of Malheur National Wildlife Refuge

  February 13: Justice Antonin Scalia dies

  June 12: Orlando nightclub shooting

  August 13: Unrest in Milwaukee

  September 26: First Presidential Debate

  October 22: Clinton testified before the House Select Committee on Benghazi

  November 8: Presidential election

  December 19: Electoral College votes

- 2017

  January 20: Presidential inauguration

  May 9: President fires FBI Director James Comey

  August 11: Unite the Right rally in Charlottesville

  December 22: Tax Cuts and Jobs Act signed into law

- 2018

  February 16: Investigation into Russian interference

  June 12: President-Kim Jong-un summit in Singapore

  October 6: Brett Kavanaugh confirmed to Supreme Court

  November 6: Midterm elections

- 2019

  January 3: 116th Congress begins (Dem Majority)

March 12: College admissions bribery scandal

April 18: Mueller Report released

August 3: Mass Shootings and Gun Control Debate

October 6: US Troop Withdrawal from Syria & Turkish Offensive

December 18: Congress impeaches President Trump (1st time)

- 2020

  February 5: Senate impeaches President Trump (1st time)

  March 13: COVID-19 national emergency declared

  May 25: Murder of George Floyd

  September 18: Justice Ruth Bader Ginsburg dies

  November 3: Presidential election

- 2021

  January 6: Capitol Hill riot

## A.10 Topic Modelling (Paper 2)

Table 7: Summary of the BERTopic outputs by dataset

| Dataset | Original N Topics | Descriptive Stats |
|---|---|---|
| Democrat Politicians | 1,500 | count: 85,809<br>mean: 0.763567<br>std: 0.274426 |
| Republican Politicians | 1,127 | count: 67,136<br>mean: 0.765548<br>std: 0.264784 |
| Dem Media | 1,654 | count: 63,957<br>mean: 0.770168<br>std: 0.285544 |
| Rep Media | 1,842 | count: 68,282<br>mean: 0.770233<br>std: 0.286560 |
| Public | 6,810 | count: 581,692<br>mean: 0.772936<br>std; 0.270832 |

*Note: The table shows the descriptive statistics referring to all topics, minus the meaningless "-1" bin, by each dataset, as well as presenting the number of all the topics found before topic reduction. The statistics refer to (1) the count of tweets to which a topic has been assigned, (2) the average of all the probabilities of the assigned topic per each tweet in the dataset, (3) probabilities' standard deviation.*

Figure 24: Topic Modelling - Democrat Politicians

**Summary of Topics (Dem Politicians)**

| Original Name | Clean New Name | Count |
|---|---|---|
| 0_trump_the_to_is | Trump and Election Discussions | 58315 |
| -1_the_to_and_trump | Trump-related Conversations | 55613 |
| 1_election_in_to_the | Election News and Updates | 13429 |
| 2_women_hillary_she_her | Women and Hillary Clinton | 5197 |
| 3_de_la_que_puerto | Puerto Rico and Related Issues | 1872 |
| 4_gun_violence_nra_the | Gun Violence and NRA Discussions | 1102 |
| 5_birthday_happy_and_you | Birthday Wishes and Celebrations | 984 |
| 6_axefiles_for_at_join | Social Media and Technology | 895 |
| 7_pompeo_nomination_blocking_stoppompeo | Mike Pompeo Nomination and Politics | 642 |
| 8_coverupcaucus_representative_part_find | Political Cover-ups and Investigations | 491 |
| 9_covering_uncovertrump_investigatetrump_wondering | Trump Investigation and Cover-ups | 481 |
| 10_smears_bots_internet_spread | Misinformation and Internet Bots | 434 |
| 11_bob_hugin_menendez_ginsburg | Political Figures: Bob Hugin and Menendez | 381 |
| 12_miners_patriotic_coal_brave | Miners and Patriotic Discussions | 272 |
| 13_massive_windfall_sick_gave | Economic Windfalls and Social Issues | 258 |
| 14_sticker_sale_store_your | Retail and Store Discussions | 222 |
| 15_notnormal_notok_flashbackfriday_normalize | Social Movements and Flashback Fridays | 174 |
| 16_gagrule_options_removes_nurses | Healthcare Issues and Nurses | 116 |
| 17_worth_moral_human_issue | Moral and Humanitarian Issues | 93 |
| 18_kkk_duke_heckno_david | Racism and Hate Groups | 92 |
| 19_developers_owners_estate_chose | Real Estate and Developer Choices | 87 |
| 20_above_remind_law_americans | Legal and American Rights Discussions | 82 |
| 21_locker_glass_room_ceiling | Gender Equality and Glass Ceiling | 70 |
| 22_launch_dogged_celebrating_truth | Truth and Media Celebrations | 52 |
| 23_right_sides_angry_wrong | Political Disagreement and Ideological Conflict | 43 |
| 24_14_attention_24_familiesbelongtogether | Immigration and Family Rights | 24 |
| 25_16m_strong_gunreformnow_fest | Gun Reform Movements | 23 |
| 26_amazon_netflix_paid_taxes | Corporate Taxes and Amazon/Netflix | 12 |
| 28_laughing_myths_explains_action | Myths and Explaining Action | 11 |
| 27_khzirkhan_dumptrump_message_watch | Trump Opposition and Media Watch | 11 |

Figure 25: Topic Modelling - Republican Politicians

**Summary of Topics (Rep Politicians)**

| Original Name | Clean New Name | Count |
|---|---|---|
| -1_the_to_and_trump | Trump and Political Discussions | 46067 |
| 0_the_trump_to_is | Trump-related Issues | 25532 |
| 1_to_the_in_president | Presidential Discussions | 17188 |
| 2_tax_the_to_jobs | Tax and Employment Discussions | 7253 |
| 3_china_iran_the_to | International Relations: China and Iran | 5587 |
| 4_biden_joe_he_hunter | Biden, Family, and Hunter Biden Discussions | 5221 |
| 5_coronavirus_the_president_drug | Coronavirus and Public Health | 1996 |
| terviewed_celebapprentice_apprentice_celebrityappre | Celebrity Apprentice Discussions | 1113 |
| 7_life_pro_abortion_religious | Abortion, Religion, and Life Discussions | 1061 |
| 8_book_you_great_thank | Book Recommendations and Gratitude | 447 |
| 9_travel_ban_flag_trump | Travel Bans and Trump's Policies | 395 |
| 10_venezuela_cuba_de_cuban | Venezuela, Cuba, and International Politics | 371 |
| 11_nfl_players_football_coach | NFL, Football, and Coaching Discussions | 194 |
| 12_entrepreneurs_yourself_your_you | Entrepreneurship and Self-help Discussions | 137 |
| 13_tea_party_uk_libertarian | Tea Party, Libertarianism, and Political Movements | 117 |
| 14_tickets_join_me_at | Event and Invitation Discussions | 107 |
| 15_teleprompter_telehealth_biden_mobile | Teleprompter, Telehealth, and Biden's Mobile Policies | 75 |
| 16_kung_marketing_materials_flu | Marketing, Flu, and Health-related Topics | 71 |
| 17_soul_sell_legacy_destruction | Legacy, Soul, and Destruction Issues | 40 |
| 18_abandoned_emails_senate_over | Political Scandals: Emails and Senate Investigations | 36 |
| 19_kanye_song_music_west | Kanye West and Music Discussions | 35 |
| 20_patrick_grow_ireland_taxes | Taxes, Ireland, and Economic Issues | 31 |
| 21_1220_349_615_walsh | Miscellaneous Political and Personal References | 25 |
| 22_clerks_diligently_throughout_county | Clerks and County Operations | 14 |
| 23_zuckerman_mort_worthless_dopey | Personal and Business Criticisms | 13 |
| 24_ایران_به_است_در | Iran and Related Political Topics | 13 |
| 25_ojeda_richard_leftist_return | Political Figures: Ojeda and Leftist Views | 13 |
| 26_scooby_van_doo_mystery | Scooby-Doo and Mystery-themed Topics | 12 |
| 27_ballots_mailed_cast_advance | Voting and Ballots Discussions | 11 |
| 28_sweden_refugees_muslim_outraged | Refugee and Muslim Issues in Sweden | 11 |

Figure 26: Topic Modelling - Democrat-leaning Media

**Summary of Topics (Dem Media)**

| Original Name | Clean New Name | Count |
|---|---|---|
| 0_election_the_to_trump | U.S. Election and Trump | 40255 |
| -1_the_election_to_in | General Election Topics | 28939 |
| 1_russian_russia_election_2016 | Russian Election Interference | 9322 |
| 2_french_election_the_france | French Election | 5682 |
| 3_election_in_presidential_of | Presidential Elections | 4603 |
| 4_netanyahu_israel_iran_israeli | Israel and Netanyahu | 1503 |
| 5_gay_abortion_lgbtq_perry | LGBTQ and Abortion | 477 |
| 6_turkey_erdogan_istanbul_turkish | Turkish Politics and Erdogan | 446 |
| 7_china_hong_kong_beijing | China and Hong Kong | 438 |
| 8_postal_usps_postmaster_service | USPS and Postal Service | 362 |
| 9_gun_nra_control_guns | Gun Control and NRA | 206 |
| 10_fame_hall_nfl_garrity | NFL and Hall of Fame | 126 |
| 11_saudi_arabia_rooms_lobbyist | Saudi Arabia and Lobbying | 60 |
| 12_bomb_racial_violence_detonate | Bombing and Racial Violence | 51 |
| 13_sell_merchandise_neil_space | Merchandise and Space Items | 48 |
| 14_asteroid_earth_day_mission | Asteroid Missions | 37 |
| 15_rmu_fired_11_au | Job Termination Cases | 31 |
| 16_safety_wearing_pins_masks | Safety Pins and Masks | 30 |
| 17_border_wall_bypassed_construction | Border Wall and Construction | 27 |
| 18_reflects_10_ideal_average | Reflection and Ideals | 25 |
| 19_fixed_thing_matter_no | Matters and Fixing Issues | 25 |
| 20_transit_strike_philadelphia_workers | Philadelphia Transit Strike | 20 |
| 21_decentralized_curse_blessing_viewed | Decentralization Discussions | 17 |
| 22_jury_trial_selection_roger | Jury Selection and Trials | 16 |
| 23_anthem_sits_evans_protest | National Anthem Protests | 15 |
| 24_human_mulvaney_terrible_being | Humanity and Criticisms | 14 |
| 25_nratv_audiences_brands_conspiracies | NRA TV and Conspiracies | 14 |
| 26_esper_defense_fires_pentagon | Pentagon and Defense Issues | 13 |
| 27_jay_cleveland_concert_headline | Cleveland Concerts and Headlines | 12 |
| 28_valdez_lupe_abbott_greg | Texas Politics (Valdez and Abbott) | 12 |

Figure 27: Topic Modelling - Republican-leaning Media

**Summary of Topics (Rep Media)**

| Original Name | Clean New Name | Count |
|---|---|---|
| 0_election_trump_the_to | Trump and Election Discussions | 51485 |
| -1_the_election_to_in | Election Campaigns and Political Discussions | 35677 |
| 1_election_french_johnson_brexit | French Elections, Brexit, and Political Movements | 5930 |
| 2_tcot_election_in_news | Conservative News and Election Coverage | 4757 |
| 3_venezuela_immigration_presidential_brazil | Venezuela, Immigration, and Brazilian Politics | 1941 |
| 4_netanyahu_israel_israeli_iran | Netanyahu, Israel, Iran, and Middle Eastern Politics | 1675 |
| 5_georgia_special_ossoff_abrams | Georgia Elections: Ossoff and Abrams | 1019 |
| 6_barr_selection_trump_parscale | Trump, Barr, and Parscale: Political Appointments | 429 |
| 7_postmaster_postal_service_changes | Postal Service and Postmaster Changes | 171 |
| 8_fireworks_cancels_lena_killed | Fireworks and Tragic Events | 145 |
| 9_apologized_apologizes_apology_ptsd | Apologies, PTSD, and Mental Health Issues | 102 |
| 10_kushner_jared_bank_kudlow | Kushner, Jared, and Economic Policy | 97 |
| 11_felons_illegally_prison_woman | Felons, Prison, and Legal Issues | 89 |
| 12_resigns_resign_ellison_after | Resignations and Political Figures | 78 |
| 13_whatsapp_oliver_elxn42_signing | WhatsApp and Election Signing Issues | 68 |
| 14_openly_transgender_milk_gay | Transgender Issues and LGBTQ+ Rights | 63 |
| 15_sidewire_livewire_hewitt_hugh | Media and Political Coverage by Sidewire and Hewitt | 57 |
| 16_asteroid_earth_miles_nasa | Asteroids and Space Exploration | 45 |
| 17_first_capitol_thisdayinhistory_presidential | Capitol and Presidential History | 39 |
| 18_farenthold_blake_duncan_abbott | Political Figures: Farenthold, Blake, Duncan, and Abbott | 33 |
| 19_native_hawaiian_halts_controversial | Native Hawaiian Issues and Controversial Policies | 23 |
| 20_état_phases_djt_ending | Donald Trump (DJT) and Political Phases | 21 |
| 21_hof_dennis_brothel_nevada | Dennis Hof, Nevada, and Brothel-Related Discussions | 20 |
| 22_mate_selection_chairs_committee | Committee Chair Selection and Political Discussions | 16 |
| 23_speeches_radio_stations_broadcast | Political Speeches and Radio Broadcasts | 16 |
| 24_ethnically_mixed_racial_districts | Racial and Ethnic Districting Issues | 15 |
| 25_invite_associate_attended_arrested | Invitations, Associations, and Arrest Incidents | 14 |
| 26_pornography_zoom_indiana_interrupted | Pornography, Zoom, and Legal Interruptions | 12 |
| 27_linkedin_apologizes_founder_disinformation | LinkedIn, Apologies, and Disinformation | 12 |
| 28_cpc_membership_buy_leadership | CPC Membership and Leadership Politics | 11 |

Figure 28: Topic Modelling - Public

**Summary of Topics (Public)**

| Original Name | Clean New Name | Count |
|---|---|---|
| 0_the_election_to_is | Election-related Discussions | 574020 |
| -1_the_election_to_and | Election and Political Discussions | 538687 |
| 1_you_it_that_to | Miscellaneous Conversations | 1820 |
| 2_oromia_prison_olf_potential | Oromia and Prison Issues | 1045 |
| 3_birthday_my_day_happy | Birthday Wishes and Celebrations | 1041 |
| 4_mccullough_newyork_william_3rd | New York and William McCullough | 836 |
| 5_followers_gained_increase_current | Social Media Growth and Followers | 414 |
| 6_watkins_qanon_ron_jim | QAnon and Related Figures | 368 |
| 7_inventor_drum_burning_jova | Inventions and Technology Discussions | 348 |
| 8_population_county_density_sq | Population and Demographic Data | 297 |
| 9_goldberg_radio_elevator_pitch | Media and Radio Discussions | 134 |
| 10_peso_lago_mar_currencies | Currency and Economic Discussions | 84 |
| 11_glasses_colored_rose_mediamafia | Fashion and Media Influence | 81 |
| 12_nato_negotiate_ineffective_ability | NATO and Global Negotiations | 77 |
| 13_mcinerney_scorecard_flying_attorneys | Legal and Political Discussions | 75 |
| 14_football_measuring_chains_game | Sports: Football and Measurement | 70 |
| 15_noun_merriam_ry_spir | Language and Vocabulary Discussions | 70 |
| 16_charged_woman_allegedly_official | Legal Issues: Alleged Charges | 68 |
| 17_sos_ss_01010101010101007_dominion | Technology and Dominions Discussions | 57 |
| 18_math_education_degree_you | Education and Mathematics | 48 |
| 19_blacklivesmatter_played_wake_realize | Black Lives Matter Movement | 46 |
| 20_public_servant_obligation_accepts | Public Service and Responsibilities | 45 |
| 21_judson_songs_memorial_recorded | Music and Memorial Discussions | 45 |
| 22_casinos_cameras_vegas_tech | Casinos, Technology, and Surveillance | 39 |
| 23_antifah_antifia_identifying_entry | Antifa and Political Identifications | 33 |
| 24_ocean_oceans_contaminating_billion | Environmental Issues: Oceans and Pollution | 25 |
| 25_panacea_scales_testimony_assume | Health and Scientific Discussions | 18 |
| 26_600_lockdown_should_dollars | Economic Impact of Lockdowns | 17 |
| 27_vikas_kaveri_inferior_pillar | Cultural and Social Issues | 11 |
| 28_1968_industrial_mcgovern_complex | Historical and Political Events | 11 |
| 1343637857076981762 | Miscellaneous ID | 1 |

## A.11 Further Details on Time-Series Annotations

The following notes refer to Figures 2 to 5 in Paper 2.

In the Democrat-coded tweets, divisiveness is higher among Republican politicians up to early 2017, suggesting that, from a Democrat-coded perspective, Republican discourse is perceived as more polarising in that period. After that, the gap between Democrats' and Republicans' levels of divisiveness narrows, with both groups displaying similar fluctuations. Significant peaks in divisiveness for Republicans coincide with events like Donald Trump's election, mass shootings and unrest (San Bernardino 2015, Orlando 2016, Milwaukee 2016), Trump's impeachment proceedings, and key legislative events such as the confirmation of Justice Kavanaugh (October 2018). For Democrats, heightened divisiveness appears prominently during moments like investigations into Russian interference in the 2016 election (early 2018), the college admissions bribery scandal (April 2019), and Trump's impeachment (December 2019). The public's divisiveness scores remain relatively stable and lower than partisan groups until early 2019, after which a noticeable rise occurs, likely influenced by increasing polarisation surrounding political events and media discourse. Noteworthy public peaks align with the occupation of the Malheur National Wildlife Refuge (January 2016), US troop withdrawals from Syria (October 2019), the Senate impeachment trial of Trump (February 2020), the murder of George Floyd (May 2020), and the Capitol riot (January 2021). Finally, media tweets exhibit relatively low and stable divisiveness over time. However, notable spikes occur for both Republican- and Democrat-leaning media accounts during the Charlottesville rally (August 2017), the US troop withdrawal from Syria (October 2019), and the murder of George Floyd (May 2020).

In the Republican-coded tweets, divisiveness scores for Democrats are consistently higher than those for Republicans, except for the period before November 2016. This pattern mirrors the trend observed in the Democrat-coded plot but with a much greater magnitude, as all divisiveness scores are nearly double. In this framing, Democrats emerge as the more divisive group among politicians, with noticeable spikes during key events such as Trump's inauguration speech (January 2017), Trump firing FBI Director James Comey (May 2017), the Charlottesville rally (August 2017), the passing of significant tax cuts (December 2017), investigations into Russian interference (early 2018), and the US troop withdrawal from Syria (October 2019). Conversely, Republican peaks appear around issues that may provoke strong conservative reactions, such as the Supreme Court legalising same-sex marriage nationwide (June 2015), mass shootings (Oregon 2015, Orlando 2016), Clinton testifying before the House Select Committee (October 2016), Trump's impeachment proceedings (2019), the death of Ruth Bader Ginsburg (September 2020), Biden's election victory (November 2020), and events leading up to the Capitol riot (January 2021). As in the Democrat-coded analysis, the public's divisiveness scores remain relatively low and stable until

early 2019, after which a sharp increase occurs, likely influenced by Twitter's "Topics" feature reinforcing content engagement loops. Significant spikes appear during major events such as unrest in Milwaukee (August 2016), the occupation of the Malheur National Wildlife Refuge (January 2016), mass shootings (August 2019), US troop withdrawal from Syria (October 2019), Trump's impeachment (December 2019 and February 2020), and the Capitol riot (January 2021). Similar to the previous analysis, media tweets are generally perceived as low and stable over time, yet with higher divisiveness scores (nearly double) compared to their Democrat-coded counterparts. Notable spikes in media divisiveness align with the occupation of the Malheur National Wildlife Refuge (January 2016), unrest in Milwaukee (August 2016), mass shootings (August 2019), and the murder of George Floyd (May 2020).

In the Combined-coded tweets, a similar pattern emerges to that observed in the Democrat-coded tweets: Republicans use highly polarising language leading up to the 2016 election, after which the levels of polarisation between Democrats and Republicans become less distinguishable overall. For Democratic politicians, notable spikes in divisiveness occur around key events such as Trump firing FBI Director James Comey (May 2017), the Charlottesville rally (August 2017), the passage of significant tax cuts (December 2017), and the investigations into Russian interference (early 2018). For Republican politicians, peaks align with issues that might provoke strong conservative reactions, including the Supreme Court legalising same-sex marriage nationwide (June 2015), mass shootings and unrest (Oregon 2015, Orlando 2016, Milwaukee 2016), Clinton testifying before the House Select Committee (October 2016), the release of the Mueller Report (April 2019), and the withdrawal of US troops from Syria (October 2019). Among the public, a strong spike corresponds to the unrest in Milwaukee (August 2016), followed by a prolonged increase in divisiveness after April 2019, similar to previous datasets. Additional peaks appear in response to mass shootings (August 2019), the US troop withdrawal from Syria (October 2019), Trump's impeachment proceedings in Congress and the Senate (December 2019 and February 2020, respectively), and the Capitol attack (January 2021). For media users, polarisation levels remain relatively low, with fluctuations staying around 10%, although small spikes seem to align with major political events.

The GPT-coded plot shows that Democrat-aligned users exhibit the highest levels of predicted divisiveness over time. However, Republican-aligned users had higher divisiveness levels from early 2015 to late 2016, after which Democrat-aligned users surpassed them significantly. This trend remained until approximately April 2019, when public discourse reached similar or higher levels, with a brief resurgence in divisiveness during the 2020 election period. It is important to note that GPT-4o identified a general increase in divisiveness across all user groups over time. For Democrat-aligned users, key spikes in divisiveness are observed around Trump firing FBI Director James Comey (May 2017), the Unite the Right rally in Charlottesville (August 2017), the release of the Mueller Report (April 2019), and the withdrawal of US troops from Syria (October 2019).

Republican-aligned users show peaks during the occupation of Malheur National Wildlife Refuge (January 2016), mass shootings and related unrest (Orlando in June 2016, Milwaukee in August 2016), Clinton's testimony before the House Select Committee (October 2016), the withdrawal of US troops from Syria (October 2019), and the election of Joe Biden (November 2020). For the public, notable peaks align with major events such as the Malheur National Wildlife Refuge occupation (January 2016), unrest in Milwaukee (August 2016), and several mass shootings and gun control debates (August 2019). Additional spikes are evident during Trump's impeachment proceedings (December 2019 and February 2020), the COVID-19 national emergency declaration (March 2020), and the Capitol riot (January 2021). Media-aligned users, particularly Democrat-leaning media, exhibit significant spikes around mass shootings and gun control debates (August 2019), the US troop withdrawal from Syria (October 2019), Trump's Senate impeachment trial (February 2020), and the Capitol riot (January 2021).

## A.12 All VAR Models

Figure 29: Democrat Coded - Cumulative Non-Orth IRF

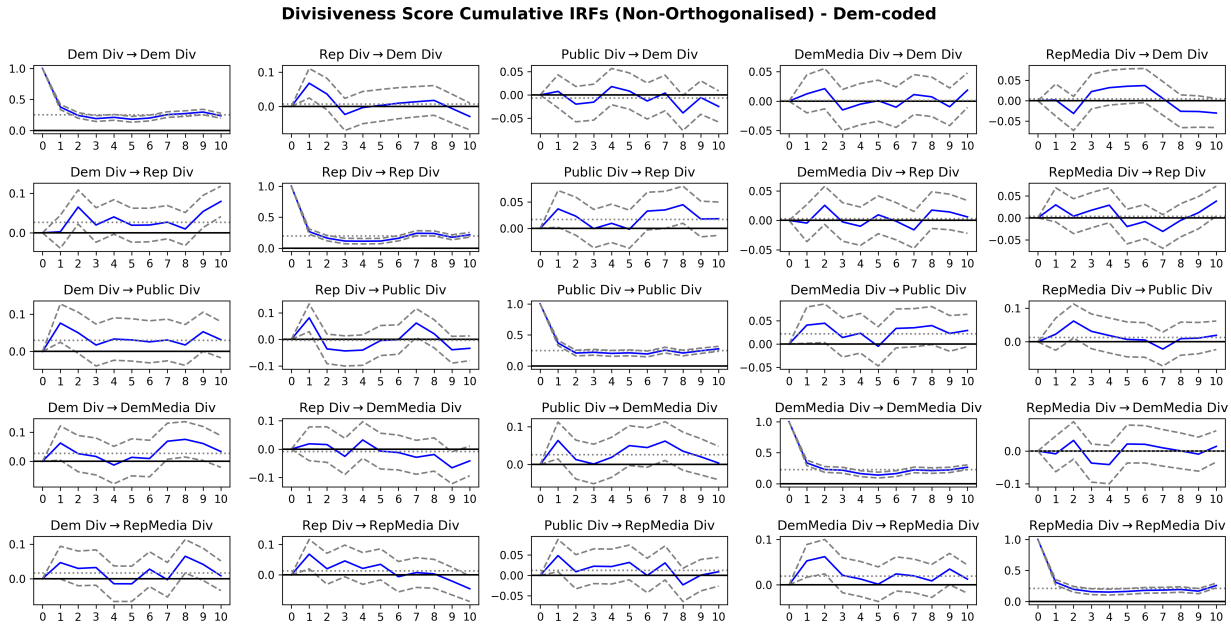



Figure 30: Democrat Coded - Cumulative Orth IRF

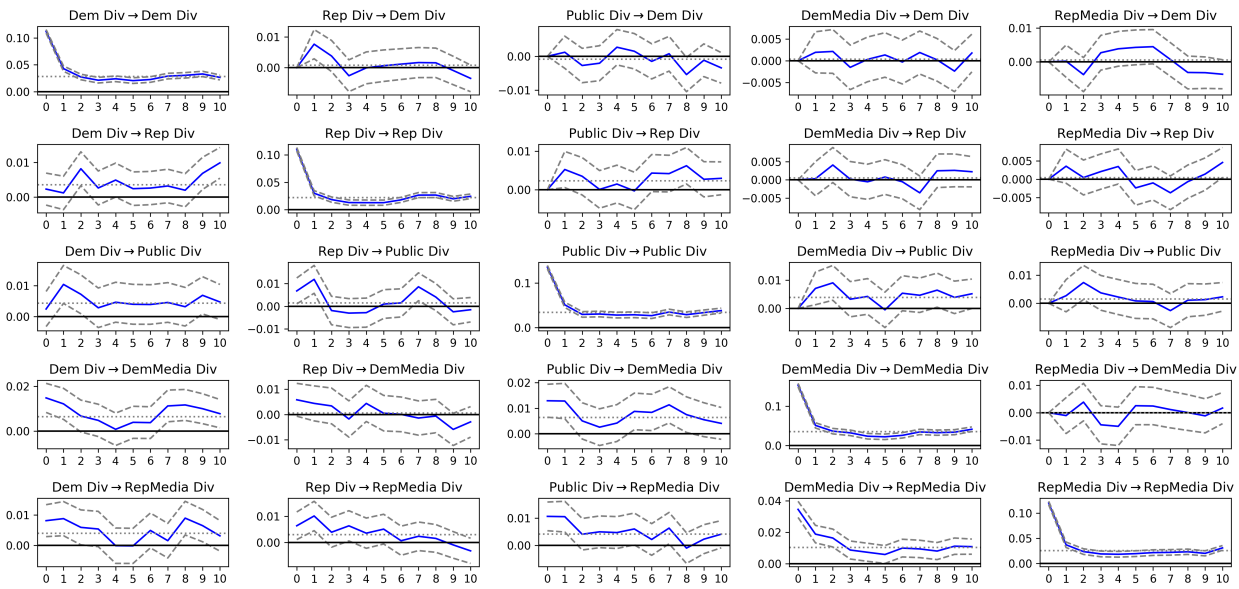


Figure 31: Democrat Coded - Non-Cumulative Orth IRF

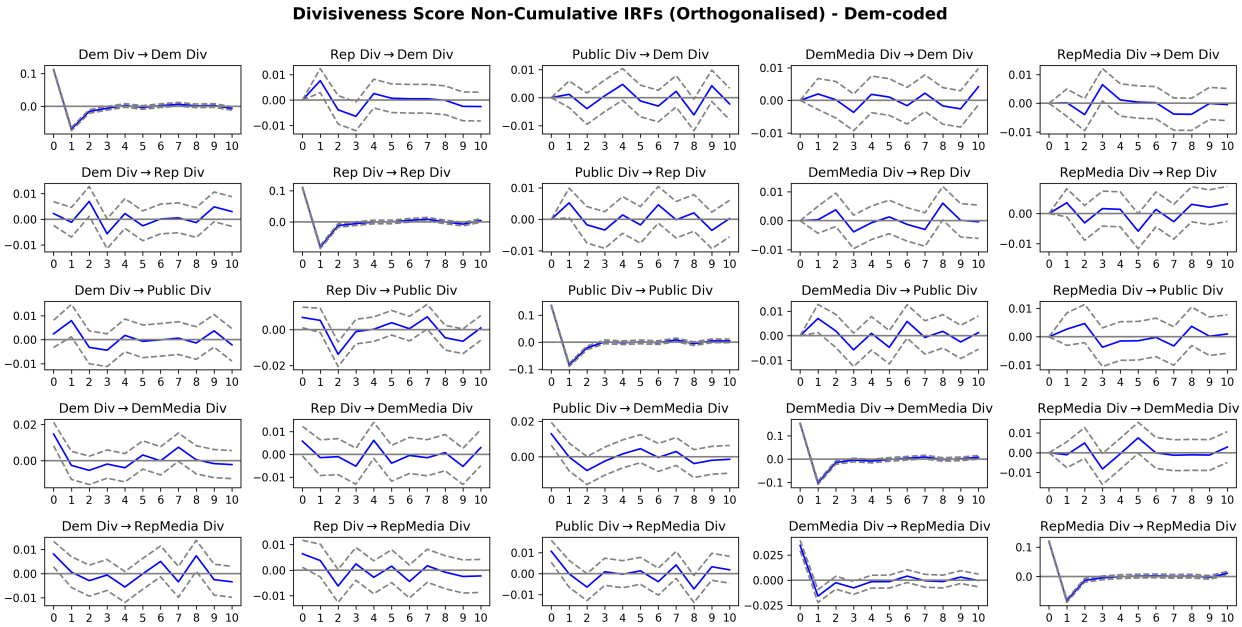



Figure 32: Republican Coded - Cumulative Non-Orth IRF

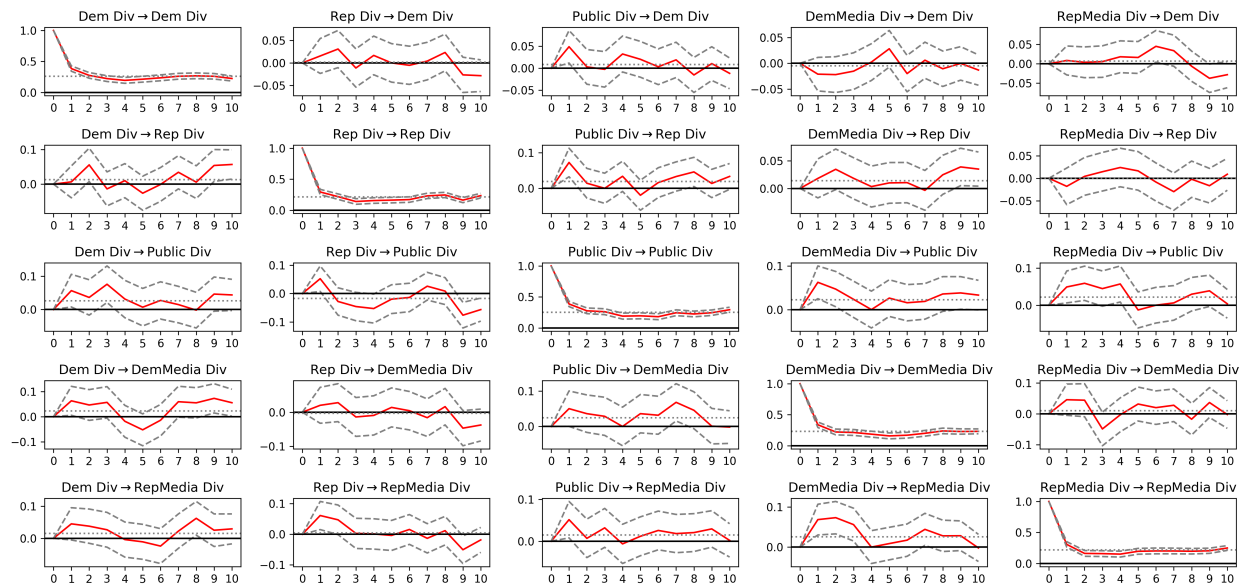


Figure 33: Republican Coded - Cumulative Orth IRF

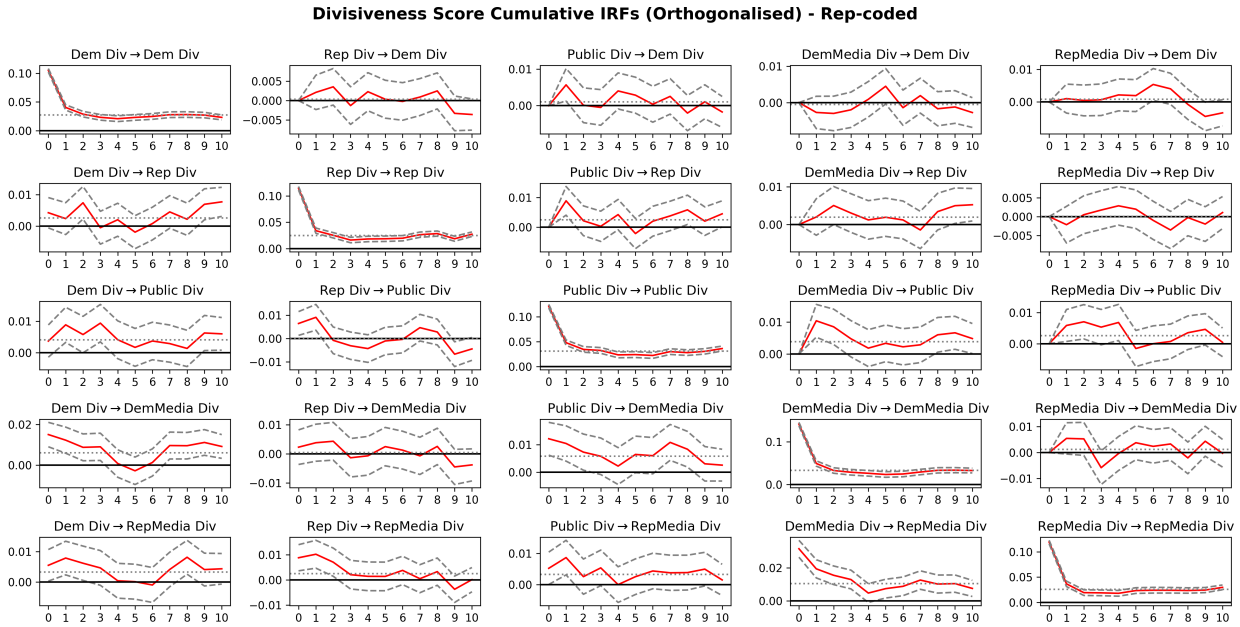



Figure 34: Republican Coded - Non-Cumulative Orth IRF

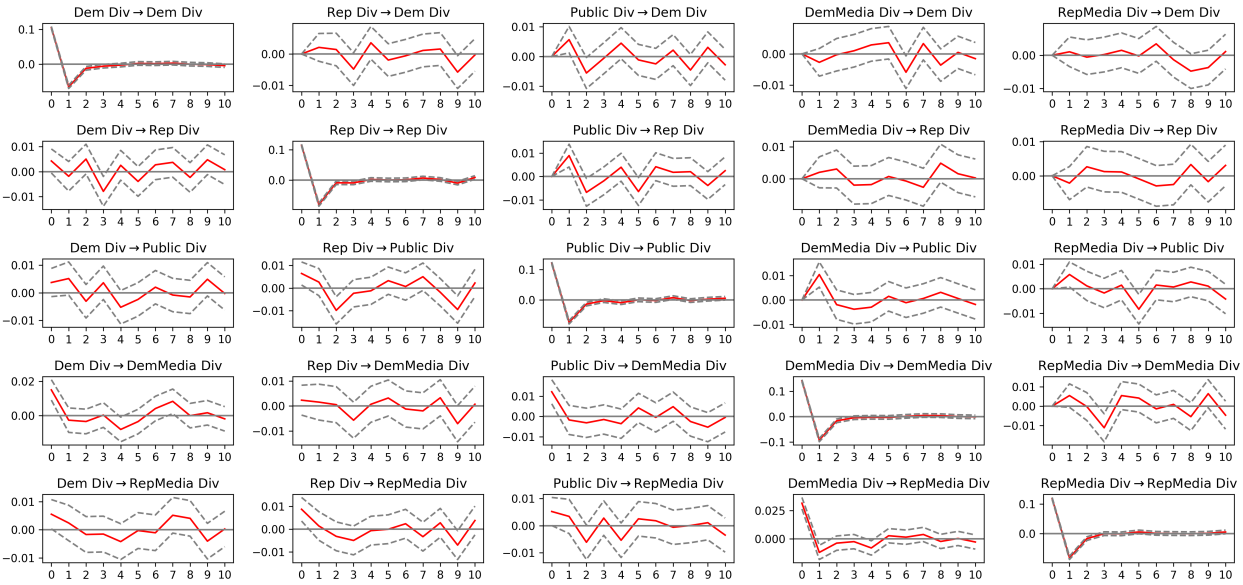


Figure 35: Combined Coded - Cumulative Non-Orth IRF

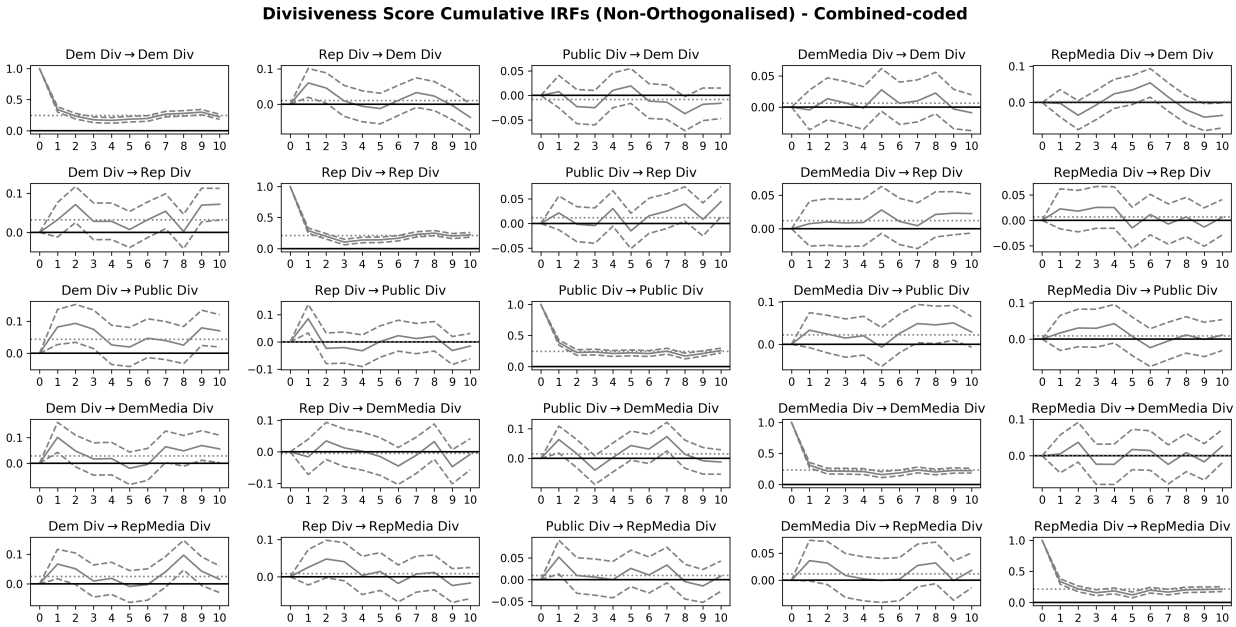



Figure 36: Combined Coded - Cumulative Orth IRF

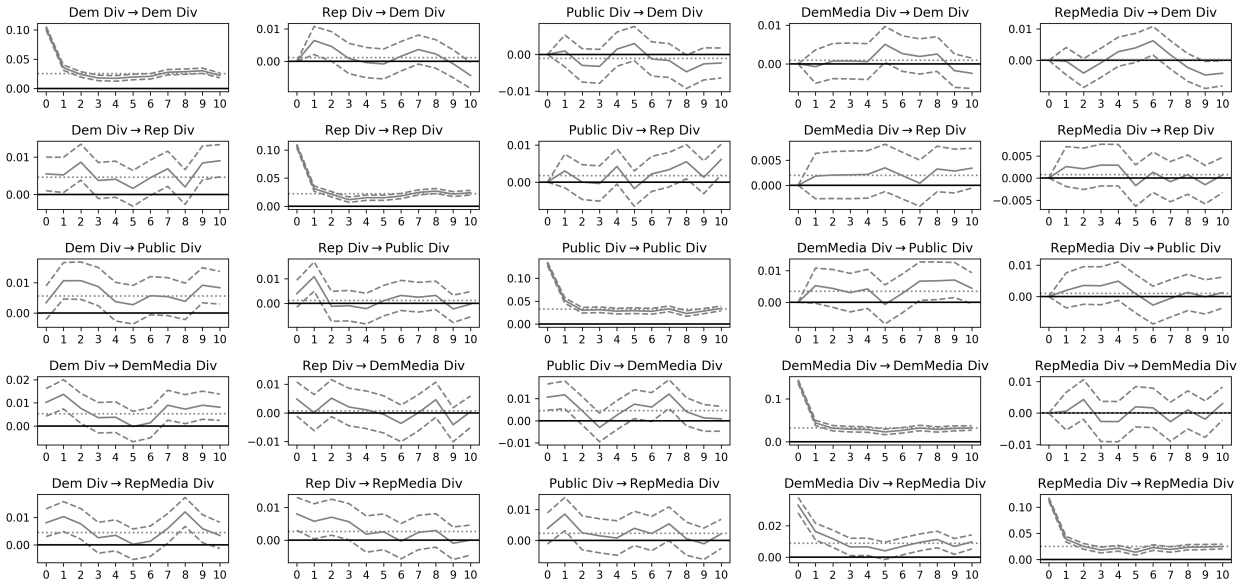


Figure 37: Combined Coded - Non-Cumulative Orth IRF

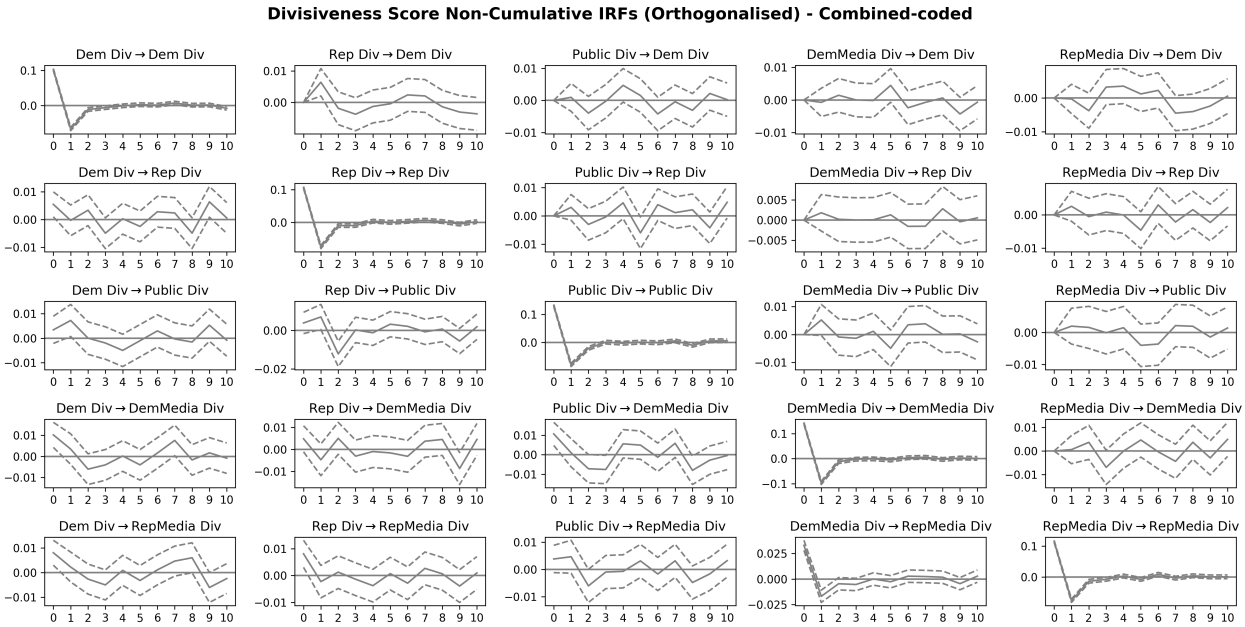



Figure 38: GPT Coded - Cumulative Non-Orth IRF

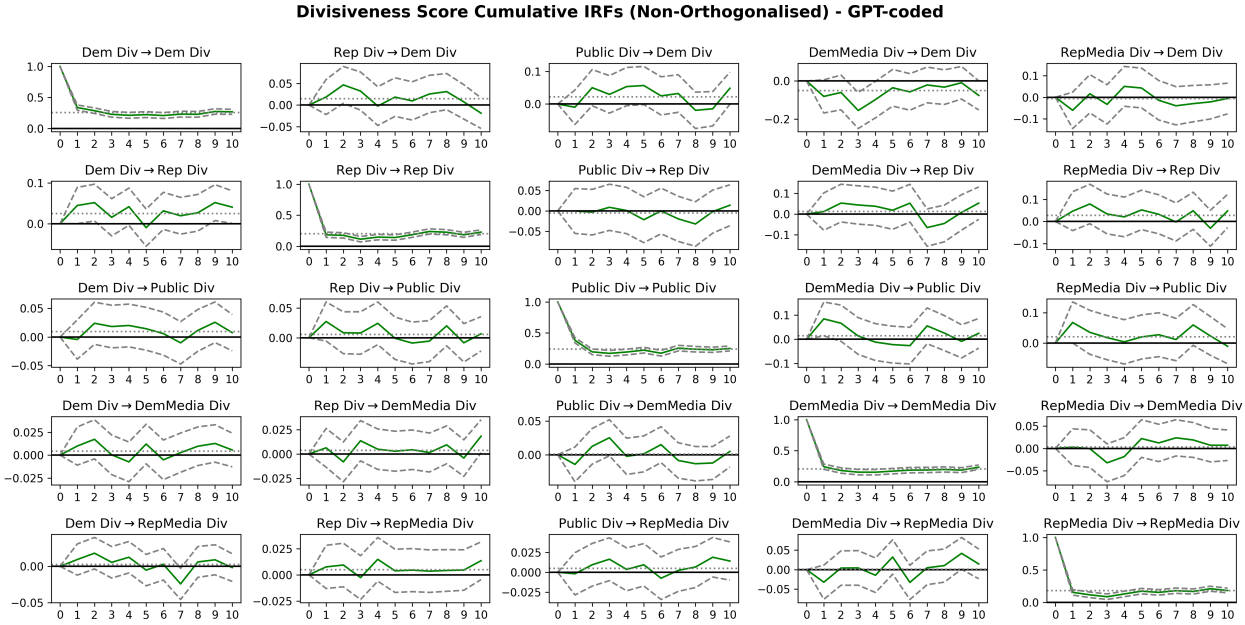


Figure 39: GPT Coded - Cumulative Orth IRF

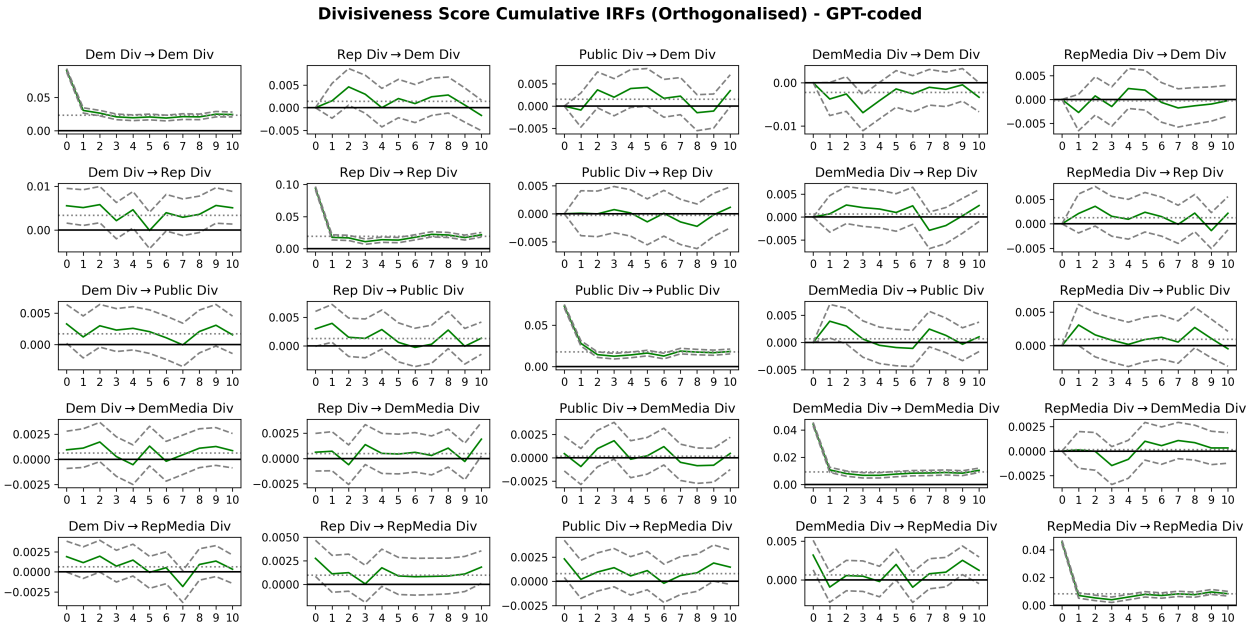
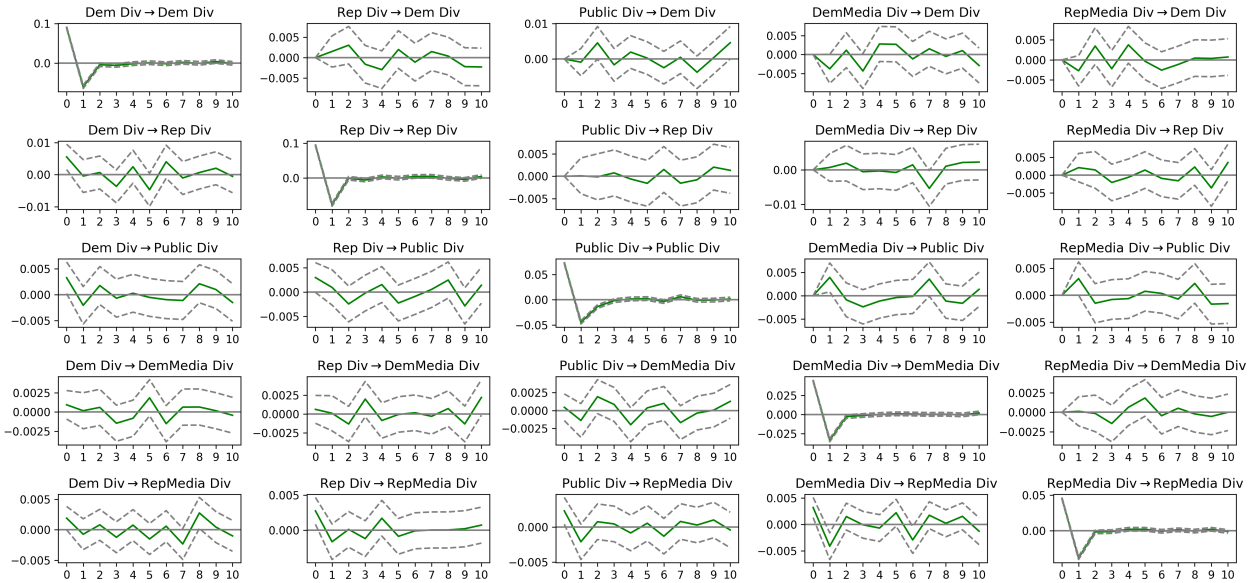



Figure 40: GPT Coded - Non-Cumulative Orth IRF



## A.13 Full HERGMs

**Democrat-coded:**

Call: bigergm(net edges + mutual + triangles + nodematch("party") + edgecov(div$_{scaled}$) + $edgecov(wgt_scaled)$+$gwesp(decay = 0.5, fixed = TRUE), n_blocks = 2, ncores = num_cores, seed =$ 42, $verbose = TRUE, control = control.ergm(MCMC.burnin = 20000, MCMC.interval =$ 500, $MCMC.samplesize = 50000$))

Found 2 clusters of relative sizes: 0.9940945 0.005905512

Results of within-cluster estimation: Maximum Pseudolikelihood Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.865972 0.037352 0 -157.046 < 1e-04 *** mutual 0.438004 0.052071 0 8.412 < 1e-04 *** triangle 0.113277 0.002053 0 55.177 < 1e-04 *** nodematch.party 0.626540 0.033560 0 18.669 < 1e-04 *** edgecov.div$_{scaled}$ − 0.2152920.0561890 − 3.8320.000127 $***$ $edgecov.wgt_scaled$ − 12.1566300.6816420 − 17.834 < $1e-04***gwesp.OTP.fixed$.0.50.5679200.020052028.323 < $1e-04***---Signif.codes$ : 0‘ $***'$ 0.001‘ $**'$0.01‘ $*'$ 0.05‘.’0.1‘’1

Results of between-cluster estimation: Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.0966 0.3188 0 -15.985 <1e-04 *** nodematch.party 0.6591 0.3854 0 1.710 0.0872 . edgecov.div$_{scaled}$1.60710.562802.8550.0043 $**$ $edgecov.wgt_scaled$ − 43.53087.00260 − 6.216 < $1e-04***---Signif.codes$ : 0‘ $***'$ 0.001‘ $**'$0.01‘ $*'$ 0.05‘.’0.1‘’1$AIC$ : 41473.32$BIC$ : 41553.2

**Republican-coded:**

Call: bigergm(net edges + mutual + triangles + nodematch("party") + edgecov(div$_{scaled}$) + $edgecov(wgt_scaled)$+$gwesp(decay = 0.5, fixed = TRUE), n_blocks = 2, ncores = num_cores, seed =$ 42, $verbose = TRUE, control = control.ergm(MCMC.burnin = 20000, MCMC.interval =$ 500, $MCMC.samplesize = 50000$))

Found 2 clusters of relative sizes: 0.9940945 0.005905512

Results of within-cluster estimation: Maximum Pseudolikelihood Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.866259 0.037356 0 -157.035 < 1e-04 *** mutual 0.439646 0.052061 0 8.445 < 1e-04 *** triangle 0.113270 0.002053 0 55.167 < 1e-04 *** nodematch.party 0.626362 0.033561 0 18.663 < 1e-04 *** edgecov.div$_{scaled}$ − 0.2264250.0584240 − 3.8760.000106 $***$ $edgecov.wgt_scaled$ − 12.1767920.6806430 − 17.890 < $1e-04***gwesp.OTP.fixed$.0.50.5681390.020054028.330 < $1e-04***---Signif.codes$ : 0‘ $***'$ 0.001‘ $**'$0.01‘ $*'$ 0.05‘.’0.1‘’1

Results of between-cluster estimation: Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.0605 0.3130 0 -16.169 < 1e-04 *** nodematch.party 0.6061 0.3811 0 1.590 0.11175 edgecov.div$_{s}caled$1.19710.445202.6890.00717 $*$ $*edgecov.wgt_{s}caled$ − 47.31315.79650 − 8.162 < $1e-04***---Signif.codes$ : 0‘ $***$’ 0.001‘ $**$’0.01‘ $*$’ 0.05‘.’0.1‘’1$AIC$ : 41487.84$BIC$ : 41567.73

**Combined-coded**

Call: bigergm(net edges + mutual + triangles + nodematch("party") + edgecov(div$_{s}caled$) + $edgecov(wgt_{s}caled)$+$gwesp(decay = 0.5, fixed = TRUE), n_{b}locks = 2, ncores = num_{c}ores, seed =$ 42, $verbose = TRUE, control = control.ergm(MCMC.burnin =$ 20000, $MCMC.interval =$ 500, $MCMC.samplesize =$ 50000))

Found 2 clusters of relative sizes: 0.992126 0.007874016

Results of within-cluster estimation: Maximum Pseudolikelihood Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.864823 0.037417 0 -156.743 < 1e-04 *** mutual 0.440720 0.052152 0 8.451 < 1e-04 *** triangle 0.113183 0.002058 0 54.997 < 1e-04 *** nodematch.party 0.632677 0.033611 0 18.823 < 1e-04 *** edgecov.div$_{s}caled$ − 0.2099790.0546490 − 3.8420.000122 $***$ $edgecov.wgt_{s}caled$ − 12.8665190.6788000 − 18.955 < $1e-04***gwesp.OTP.fixed$.0.50.5635710.020084028.061 < $1e-04***---Signif.codes$ : 0‘ $***$’ 0.001‘ $**$’0.01‘ $*$’ 0.05‘.’0.1‘’1

Results of between-cluster estimation: Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.1577 0.2848 0 -18.113 < 1e-04 *** nodematch.party 0.4561 0.3550 0 1.285 0.198875 edgecov.div$_{s}caled$1.52020.455603.3370.000847 $***$ $edgecov.wgt_{s}caled$ − 57.13675.32260 − 10.735 < $1e-04***---Signif.codes$ : 0‘ $***$’ 0.001‘ $**$’0.01‘ $*$’ 0.05‘.’0.1‘’1$AIC$ : 41408.85$BIC$ : 41489.68

**GPT-4o-coded:**

Call: bigergm(net edges + mutual + triangles + nodematch("party") + edgecov(div$_{s}caled$) + $edgecov(wgt_{s}caled)$+$gwesp(decay = 0.3, fixed = TRUE), n_{b}locks = 2, ncores = num_{c}ores, seed =$ 42, $verbose = TRUE, control = control.ergm(MCMC.burnin =$ 20000, $MCMC.interval =$ 500, $MCMC.samplesize =$ 50000))

Found 2 clusters of relative sizes: 0.9940945 0.005905512

Results of within-cluster estimation: Maximum Pseudolikelihood Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.838588 0.037405 0 -156.089 <1e-04 *** mutual 0.485188 0.051798 0 9.367 <1e-04 *** triangle 0.123114 0.001956 0 62.935 <1e-04 *** node-

match.party 0.650536 0.033511 0 19.413 <1e-04 *** edgecov.div$_scaled$ − 0.1804060.0621570 − 2.9020.0037∗∗$edgecov.wgt_scaled$−12.3532690.6786290−18.203 < 1$e$−04∗∗∗$gwesp.OTP.fixed$.0.30.55598 1$e$ − 04 ∗ ∗ ∗ − − −$Signif.codes$ : 0‘ ∗ ∗ ∗′ 0.001‘ ∗ ∗′0.01‘ ∗′ 0.05‘.′0.1‘′1

Results of between-cluster estimation: Results:

Estimate Std. Error MCMC % z value Pr(>|z|) edges -5.0524 0.3109 0 -16.252 <1e-04 *** nodematch.party 0.5924 0.3786 0 1.565 0.1176 edgecov.div$_scaled$1.02700.487702.1060.0352 ∗ $edgecov.wgt_scaled$ − 49.59245.61790 − 8.828 < 1$e$ − 04 ∗ ∗ ∗ − − −$Signif.codes$ : 0‘ ∗ ∗ ∗′ 0.001‘ ∗ ∗′0.01‘ ∗′ 0.05‘.′0.1‘′1$AIC$ : 41639.84$BIC$ : 41719.73