Investigating the Cross-Platform Behaviours of Online Hate Groups

Fatima Zahrah

Exeter College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

This thesis is dedicated to
My dear parents
For their endless love, support and encouragement.

Acknowledgements

Completion of this doctorate would not be possible without the guidance and support of several individuals.

Firstly, I would like to express my deepest gratitude to my supervisors, Dr Jason R. C. Nurse and Professor Michael Goldsmith. Their continuous support, patience and guidance were integral to the completion of this academic journey.

My sincere thanks to Professor George Ghinea for agreeing to be my external examiner. Thanks also to Dr Helena Webb and Dr Joss Wright for their insightful comments, suggestions and encouragement during my Transfer and Confirmation of Status examinations. Their expertise and constructive feedback have been invaluable in improving the quality of my research.

I would like to thank my peers at the CDT in Cyber Security. Their friendship and camaraderie have made my DPhil journey a more enjoyable and enriching experience.

I am deeply grateful to my family for your direction and prayers. Thank you to my mother and father for your love, encouragement, and guidance. Your unwavering belief in me and my abilities has been the biggest source of strength and inspiration throughout my academic journey. I would not be where I am today without you. To my sisters, thank you for tolerating my complaints and for providing me with a welcome distraction whenever I needed a break. I am also grateful to my husband for his continuous care and support.

To my dear friends, thank you for all the fun times, long conversations and for providing motivation. I'd like to especially thank Anjuli, Iffat, Hannah, Hira and Komal for their friendship and for lifting my spirits whenever I needed it.

To everyone who has contributed in many ways to this journey and made it an unforgettable experience, I am forever indebted.

Abstract

The past few decades have established how digital technologies and platforms have provided an effective medium for spreading hateful content. Despite efforts from law-enforcement agencies and platform developers to remove or limit such content, online hate ideologies and extremist narratives are still being linked to several catastrophic consequences around the world. The concept of online hate is still considered a complex phenomenon, with its definition evolving across several theoretical paradigms and disciplines, and spanning multiple forms of victimisation. Due to this complexity, research into online hate is fragmented throughout numerous disciplines, including computational social science. Previous research has demonstrated how online hate thrives globally through self-organised, scalable clusters that interconnect to form robust networks spread across multiple social-media platforms, countries, and languages. Although several extensive approaches and methods have been proposed in previous studies for the analysis of online hate, limited research has investigated how hateful behaviours and content compare and relate across different online platforms.

This thesis aimed to address these limitations by developing a cross-platform analysis framework for online-hate researchers to gain a clearer understanding of the dynamics of the global hate ecosystem. More specifically, the designing of this framework involved examining the main functionalities of existing online-hate analysis frameworks, and the extent to which they address cross-platform hate. The strengths and limitations of these approaches then informed the functional requirements of the cross-platform analysis framework. To demonstrate how the framework can provide novel insights into online-hate research, this thesis also details its application to various case studies, including online hate from white-supremacy-supporting users and environments spread during the 2020 US election and the COVID-19 pandemic.

This comprises a comparative analysis of hateful content in terms of the major topics of discussion and psycho-linguistic properties across different types of online platforms using natural language processing techniques. Additionally, the framework is used to explore networks of shared content, particularly through the posting of URLs, by harnessing social-network analysis methods. Finally, the cross-platform analysis framework is validated using a list of validation criteria to evaluate its

practicality in investigating hateful content and providing novel insights into the field of online hate. The findings from this can be used to develop more effective analysis tools for online-hate researchers and law-enforcement agencies.

Contents

List of Figures

List of Abbreviations

NLP . . . . . Natural Language Processing.

SNA . . . . . Social Network Analysis.

CSS . . . . . Computational Social Science.

ILGA . . . . . International Lesbian, Gay, Bisexual, Trans and Intersex Association.

CERD . . . . . International Convention on the Elimination of All Forms of Racial Discrimination.

API . . . . . Application Programming Interface.

TF-IDF . . . . Term Frequency-Inverse Document Frequency.

BOW . . . . . Bag of Words.

SPLC . . . . . Southern Poverty Law Center.

LDA . . . . . Latent Dirichlet Allocation.

NMF . . . . . Non-Negative Matrix Factorization.

NLTK . . . . . Natural Language Toolkit.

LIWC . . . . . Linguistic Inquiry and Word Count.

BLM . . . . . Black Lives Matter.

1 Introduction

1.1 Background and Motivation

The past few decades have demonstrated how the Internet is playing an ever-increasing role in daily life and has become an integral asset in society with a plethora of opportunities for the general public. In particular, the use of various digital technologies and online platforms for communication has been rapidly adopted into the home and workplace alike, which has been especially emphasised during the recent COVID-19 pandemic. However, this has also had several implications as various malicious actors or criminal groups are quickly exploiting both the benefits afforded by such technologies as well as the vulnerabilities presented by them for their own criminal gains. Digital technologies and platforms have provided an effective medium for spreading hateful content, and thus bring new challenges for agencies responsible for ensuring the boundaries of acceptable and legal behaviour are not crossed [1].

Though conducted in the virtual world, online hate has still affected both individuals and populations offline, with ethnic or religious minorities, people with disabilities and the LGBTQ community being predominantly targeted and influenced [2]. Spreading hateful content online has also emerged as a tool for politically motivated bigotry, xenophobia, homophobia, religious discrimination,

and excessive nationalism [3–6]. One major instance in which this has been exhibited was during the 2016 US presidential elections; the narrative of the "Make America Great Again" campaign slogan provided new possibilities for radical nationalist groups and extremist organisations, including neo-Nazis and white supremacists, to distribute their content more easily and communicate with their audiences at a much larger scale, all while granting a perceived sense of anonymity and decentralisation [2, 7]. Moreover, these online platforms have given hate-driven extremist groups a medium for launching propaganda to radicalise audiences globally [8].

Online hate has thus been linked to several abhorrent real-world events including a current surge in offline hate crimes [9]; an increase in suicides resulting from social-media vitriol [10]; inciting mass shootings, stabbings and bombings, such as the 2019 terrorist attack in Christchurch [11]; as well as the recruitment of extremists [8]. These events highlight the very real negative impacts of spreading online hate, therefore research into this field is still essential to negate its effects. Nevertheless, the concept of online hate is still considered a complex phenomenon, with its definition evolving across several theoretical paradigms, disciplines and spanning multiple forms of victimisation [12].

Due to this complexity, research into online hate is fragmented throughout numerous disciplines. Since the adverse effects of online hate are more widely recognised in society, these disciplines, including computational social science (CSS), introduce their own approaches to study and solve the associated problems. For instance, computational approaches will range from large-scale data collection [13–15] and automatic hate detection [16–19], to content diffusion analysis [20, 21] as well as the analysis of online hate networks [11, 22].

Despite all these extensive approaches and methods proposed to analyse online hate, limited research has investigated how hateful behaviours and content compare and relate across different online platforms. It has only recently been recognised within academic literature that online hate is not simply an issue for a select few platforms, rather networks of hate are often linked across these platforms, forming a global "network-of-networks" dynamic [11]. In other words, online hate thrives

globally through self-organised, scalable clusters that interconnect to form resilient networks spread across multiple social-media platforms, countries and languages. These networks formed by hate groups have proven to be remarkably resilient, and have increasingly shown to migrate across various platforms, maintaining and oftentimes expanding their connections in the process. This interconnection of several hate clusters allows for the rapid rewiring and self-repairing of the network at the micro-level when it is attacked [11].

Though there is sufficient evidence that suggests and proves such strategic usage of multiple platforms by hate groups, minimal research has been carried out to explore this further. Applying a cross-platform approach to the study of online hate would therefore be an effective way to address this gap by advancing and validating existing findings. Additionally, content moderation carried out independently on single platforms has led to less-policed platforms becoming isolated, allowing online hate to flourish. Insights gained from such research would thus help inform how platform policing and regulation can become more uniform or coordinated across multiple platforms. Current studies applying this approach for analysis are very limited, and generally provide preliminary insights.

This thesis aims to build on this particular line of research by developing a cross-platform analysis framework for the purpose of gaining a clearer understanding of the dynamics of the global hate ecosystem. In particular, this framework will make use of data retrieved from a variety of different platforms and analyse the wider structural dynamics of online hate across them. The framework is designed for CSS researchers, so as to provide them with novel insights into cross-platform online hate through more effective approaches for analysis.

1.2 Problem Statement, Research Questions and Objectives

Online hate is a very complex and topical issue fragmented across several disciplines [23–25]. When considering the fact that it is not an issue for a select few platforms, but rather it encompasses a large range of messaging and social-media platforms,

both mainstream and fringe, the challenge of understanding the impacts of online hate expands further to recognise how each of these platforms is used by hateful groups and individuals [14, 26–28]. This is especially true with regard to understanding how each specific platform contributes to a global 'network-of-networks' dynamic, and the strategic use of multiple forums to increase the resilience of online hate organisations [11, 29].

To address this problem, this research is conducted with particular consideration to the following research questions:

— **RQ1**: What are the main features and functionalities of current online-hate analysis frameworks and tools, and how do they deal with emerging research themes?

— **RQ2**: How can existing online-hate analysis frameworks be further improved in order to enhance analysis efforts of online-hate researchers and address major gaps in the current research landscape?

— **RQ3**: What insights and understanding can be derived by analysing hateful content and activities across multiple online platforms?

More specifically, the objectives of this research are:

1. **Objective 1:** To review the current research landscape in the field of online hate and identify areas for further study within academic literature. Research from various disciplines is explored, though this particularly focuses on studies and methods from CSS.

2. **Objective 2:** To understand the extent to which hate groups or hate ideologies make use of multiple platforms in spreading hateful content, and to examine whether this has been addressed in existing academic literature and associated analysis tools or frameworks. In particular, this involves a comparison of existing analysis frameworks for online hate with regards to their functionalities, the platforms they are concerned with, and any gaps identified in their analysis methods.

3. **Objective 3:** To develop a framework to aid online-hate researchers in the analysis of online hate using multiple platforms. This framework addresses and implements the requirements identified from existing approaches in the literature in Objective 2. This framework includes functionalities that provide novel insight into the cross-platform behaviours of hateful users.

4. **Objective 4:** Using the proposed framework, to gain a more comprehensive understanding of how hate groups or individuals from hateful ideologies adapt and modify their content and behaviours across different online platforms. Previous studies have shown that, although mainstream platforms have made efforts to remove hateful and extremist accounts, hate groups have tailored themselves to these policies by strategically utilising and linking their content across mainstream and fringe platforms, where there are more lax policies [30]. For instance, Hine et al. show how 4chan users "raid" YouTube videos by posting large numbers of abusive comments in a relatively small period of time [31]. Understanding how these behaviours and content change across the multiple platforms utilised by hate groups helps to recognise strategies of organised hate in the wider hate ecosystem.

5. **Objective 5:** To conduct an in-depth analysis of how multiple online platforms are strategically used to form more resilient online-hate networks, and to explore how content is shared and diffused across multiple platforms. Previous research from Johnson et al. suggests that hate organisations form a global 'network-of-networks' dynamic, where clusters of hateful users are interconnected across different platforms, enabling rapid self-repairing when blocked by platform policies [11]. Modelling the wider online hate networks thus provides a more comprehensive representation of the global hate ecosystem. In addition to this, although content diffusion dynamics have been investigated in other fields such as fake news [32, 33] and public policy debates [34], little work has been done to apply similar methodologies to online hate. This allows for a better understanding of how content moves

across different platforms and how this relates to the ways in which content is adapted for different environments.

6. **Objective 6:** To use the insights drawn from the application of the proposed cross-platform analysis framework to reflect on its validity and effectiveness in investigating hateful content. These reflections are then be used to make conclusions as to whether adopting a cross-platform approach to the research of online hate provides a more comprehensive understanding of the issue as compared to a single-platform approach, and a more extensive representation of the wider hate ecosystem.

1.3 Scope

The cross-platform analysis framework for online hate proposed in this thesis is intended to be fairly broad in research focus and use. It aims to provide a comprehensive view of the usage of multiple online platforms through various different hateful ideologies and groups, as well as different types of platforms. As the majority of the research in online-hate literature is English-language centric, this framework has been applied to English-only content throughout this thesis. Although this restricts the insights gained from the analysis framework, since most research conducted in this field is focused on English-language content, it is reasonable to assume that this framework, and any subsequent analysis tools developed, would provide valuable insight to most online-hate researchers.

In order to properly validate the proposed framework, it would ideally be applied to several other case studies, other than those used within this research. In addition to this, though the validity of the framework is discussed, it is difficult to reflect on the usability of the framework without involving other online-hate researchers willing to make use of the framework and provide supplementary feedback. Due to the magnitude of such a task and the scope of this thesis, the data collected and case studies were limited, which further narrowed the insights that could be drawn regarding the validity and effectiveness of the proposed framework. Consequently,

the core of this research is narrowed in scope to focus on specific, more informative applications of the framework, with any validation carried out concentrated on insights drawn from this specific application of the framework.

1.4 Research Contributions

The main contributions of this thesis are provided below:

1. An extensive literature review of academic research within the field of online hate. Here, definitions of online hate and other related concepts from previous research are refined. This review also highlights current trends in online hate analysis, as well as research gaps and limitations. Additionally, a comprehensive comparison of existing analysis frameworks within this field is provided, so as to determine which analysis techniques would be most insightful within the proposed framework.

2. A novel approach to studying online hate through the development of a cross-platform analysis framework. This framework makes use of multiple platforms to provide a more comprehensive and realistic understanding of the wider behaviours and narratives of hateful users. More specifically, this framework harnesses a variety of analysis techniques, including Natural Language Processing (NLP), machine learning, sentiment analysis, and Social Network Analysis (SNA) methods, to explore data collected from multiple data sources.

3. An application of the cross-platform analysis framework to gain insight into the posting behaviours and narratives promoted by hateful users on four different platforms (Twitter, Reddit, 4chan and Stormfront). This includes a thorough comparison of the posting frequency, topics of discussion, and psycho-linguistic composition of posts from these platforms, across two different case studies (the 2020 US election and the COVID-19 pandemic).

4. An application of the cross-platform analysis to identify and explore networks of shared content across the same four platforms. More specifically, this harnesses network-analysis techniques to investigate content-sharing practices through the usage of URLs on each platform, and observe the content diffusion and dissemination dynamics through cross-platform hate networks.

5. A systematic validation of the proposed cross-platform analysis framework to assess its strengths and limitations, with particular regards to its ability to provide novel understanding into the field of online hate. This includes evaluating its applicability to various case studies, as well as demonstrating how it can be used by online-hate researchers in practice.

Through these contributions, this thesis provides unique and novel perspectives into research on online hate. Additionally, the proposed cross-platform analysis framework for online hate speech demonstrates how research can benefit from perspectives and collaboration across various disciplines and domains, thus leading to innovative solutions for addressing various social issues. More specifically, the impacts and novelty it provides lie in its ability to standardise approaches, foster multidisciplinary collaboration, and adapt insights across different domains.

1.5 Thesis Outline

To address the above described research objectives, the structure of the subsequent chapters within this thesis is as follows:

Chapter 2 introduces the necessary background related to online hate by defining relevant concepts and taxonomies of online harm. A comprehensive survey of the research landscape within this field, including a comparison of current analysis tools and frameworks, is then provided to identify gaps within relevant literature, providing motivation for this thesis.

Chapter 3 presents an overview of the general research methodology used when conducting this research. Here the mixed methods from CSS that were adopted to address the objectives – thereby answering the core research questions – are discussed, as well as the ethical implications of this research.

Chapter 4 provides a detailed overview of the proposed cross-platform analysis framework for online hate. Much of this is based on the findings gained from surveying previous literature and analysis tools in Chapter 2, so as to address current research gaps. Both the conceptual basis as well as the operational structure of the framework are discussed in this chapter. This framework constitutes the main contribution of this thesis.

Chapter 5 provides a comprehensive example of how the cross-platform analysis framework detailed in the previous chapter can be applied in practice. Specifically, the focus of this chapter is to determine how different online platforms are utilised in online hate through using various content analysis techniques, including sentiment analysis and topic modelling, with particular regards to the type of content posted to each platform. Here the analysis explores online hate through two case studies: the 2020 US election and the COVID-19 pandemic.

Chapter 6 then applies other functional components of the cross-platform analysis framework to hateful content during the 2020 US election and COVID-19 pandemic. In particular, this chapter works to identify networks of hate over multiple online platforms by utilising network analysis techniques to understand how content, including URL domains, is shared across platforms. This chapter further explores the content diffusion dynamics throughout these multi-platform networks.

Chapter 7 discusses the validity of the cross-platform analysis framework by reflecting on the insights gained from applying the framework in practice across various case studies. The key findings from the previous chapters are evaluated and assessed with regards to whether using a cross-platform approach provides any meaningful insights in comparison to an analysis approach developed for a single platform.

Chapter 8 concludes the thesis and summarises the main contributions of the research project. This chapter also discusses the limitations of the thesis, most of which were identified through the validation process. Finally, it outlines future research directions that could both improve and further the ideas embodied in this work.

Figure 1.1 illustrates the structure of this thesis and the relationship between the core research questions, objectives and contributions.

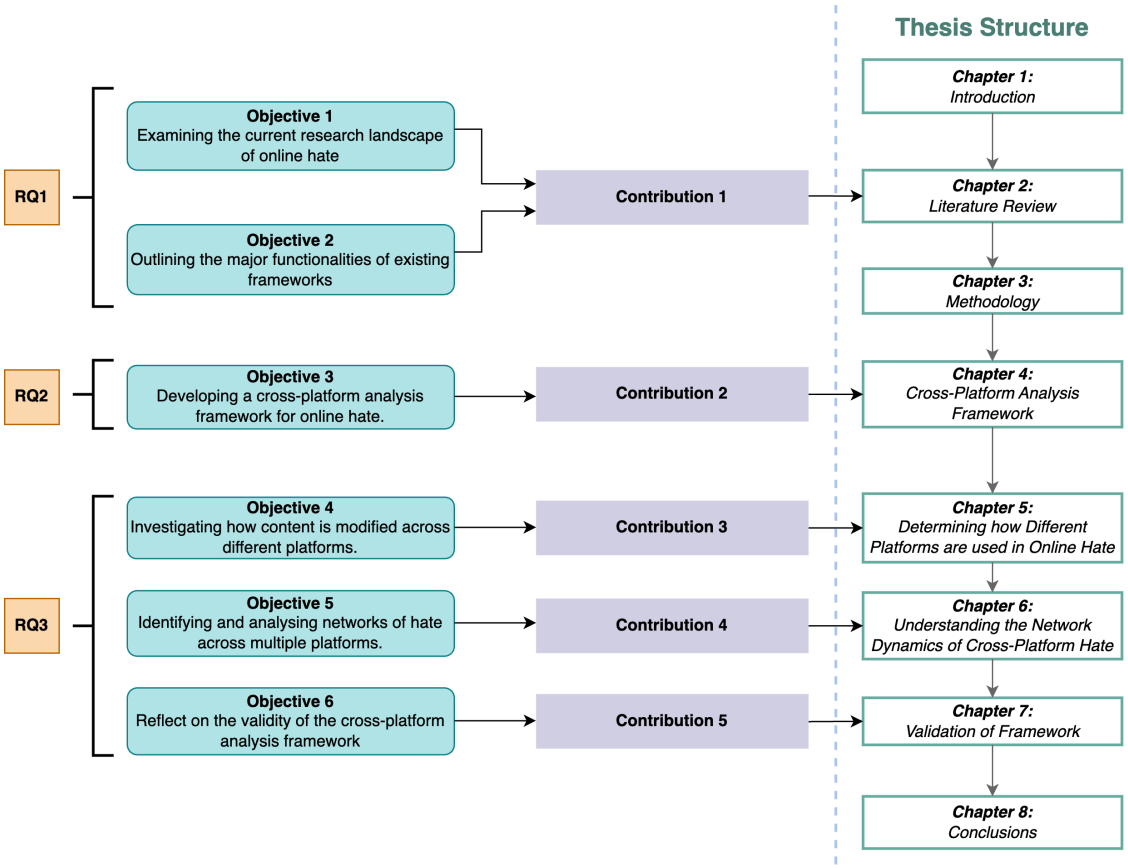

Figure 1.1: An outline of the thesis and its research contributions.

2
Literature Review

2.1 Introduction

Following the establishment of the motivations and aims of this project in Chapter 1, Chapter 2 provides the theoretical foundation for this research by conducting a thorough literature review. This review critically assesses current academic approaches and findings for the problem of online hate, with the goal of identifying outstanding gaps within the research landscape, as well as current practices in developing analysis frameworks for the study of social-media data.

Seeing as the subject of online hate has been researched extensively across a range of different disciplines, this chapter begins with defining the concept of online hate, which will then be used throughout the rest of this thesis, by examining definitions used by different agencies. This allows for the discussion of how these definitions are used to form rules utilised in identifying hateful content, and other concepts related to online hate that may be considered in the surveyed literature and the research conducted in this project.

Building upon this basis of understanding, the approaches and methodologies used across Computational Social Science (CSS) to investigate this field are then detailed. In particular, this will emphasises on the computational approaches that are most commonly used within online hate research. This involves a comprehensive

analysis of the current methods, techniques and resources used or developed to address online hate. This literature survey thus enables the identification of aspects of online hate that remain unresolved. These research gaps form the main motivations for this project, with particular regard to using a cross-platform approach to explore online hate. The findings from this literature review are then consequently used to define the requirements used in the development of the cross-platform analysis framework proposed in subsequent chapters of this thesis.

2.2 What is Online Hate?

Deciding whether certain content is 'hateful' is not simple, even for humans. Online hate is a complex phenomenon, intrinsically associated to relationships between groups, and also relying in language nuances [23]. It is oftentimes the case that there is low agreement found between annotators in the process of building new models [35, 36]. Therefore, it is crucial to clearly define online hate to make the task of its identification and analysis easier.

2.2.1 Definitions from Other Sources

In this section, different definitions of online hate are collected and perspectives from diverse sources are compared, as presented in Table 2.1. Concerning the sources of the definitions, a wide range of origins were chosen, the motivations for which are as follows:

- The European Union Commission, which is the regulator of other legislative and policy-making institutions, including governments, within the European Union.
- International minorities associations and human rights conventions, including the International Lesbian, Gay, Bisexual, Trans and Intersex Association (ILGA) and the International Convention on the Elimination of All Forms of Racial Discrimination (CERD), aim to protect people that are usually targets of online hate.

- Academic papers, to include also a perspective from academic research. Only one definition is provided in the table, though this, or definitions very similar to this, are adopted by several papers, some of which are cited in Table 2.1.
- Platform terms and conditions (including Facebook, YouTube, and Twitter), as these are commonly used to spread hate, and also work to proactively remove or limit online hate.

Table 2.1: A compilation of various definitions of online hate from several sources.

Source	**Definition**
EU Commission	"All conduct that publicly incites violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin." [37]
CERD	Any form of discrimination "based on race, colour, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life." [38]
ILGA	"Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility towards a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups." [39]
Academic Papers [40–44]	"Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity." [40]
Facebook	"A direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status. We define 'attack' as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation. We allow humour and social commentary related to these topics." [45]
YouTube	"Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity." [46]
Twitter	"Any content that promotes violence against or directly attacks or threatens other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. " [47]

To better understand the various definitions from these sources, four distinct factors are considered by which the definitions can be compared. These factors

have also been used by Fortuna and Nunes to compare definitions of hate speech within their study [23], and are detailed as follows:

- **Online hate has specific targets.** All the quoted definitions point out that online hate has specific targets and it is based on specific characteristics of groups, like ethnic origin or religion to name a couple.
- **Online hate is to incite violence or hate.** Each of the definitions use slightly different terms to describe when online hate occurs. The majority of the definitions point out that online hate is intended to incite violence or hate toward a minority (definitions from Code of Conduct, ILGA, YouTube, Twitter).
- **Online hate is to attack or diminish.** Additionally, some other definitions state that online hate is to use language or other content that attacks or diminishes these groups (definitions from Facebook, YouTube, Twitter).
- **Humour has a specific status.** However, platforms like Facebook also acknowledges that some offensive and humorous content is allowed. The exceptional status of humour makes the boundaries about what is forbidden in the platform more ambiguous.

Despite the similarities between the definitions, it can be concluded that there are some nuances that distinguish them (e.g., the usage of humour). This has provided a more comprehensive definition of online hate, which is used throughout the rest of this research.

2.2.2 Definition Used in This Thesis

In the previous section, four dimensions were used to compare various definitions of online hate. Through this analysis, a resulting definition of online hate is proposed through combining some of the key elements highlighted in the previous definitions:

> *Online hate is discourse in online environments that attacks or diminishes, or incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, or gender. It can occur through different mediums (textual or graphic) or linguistic styles, even in subtle forms or when humour is used.*

Though this definition is similar to those from other sources, it is distinct with regards to incorporating how more understated forms of hate, such as humour, can also be offensive, which has only recently been acknowledged in literature and definitions [48]. The occurrence of violence can manifest in both explicit physical forms as well as subtle ways. This is particularly evident when stereotypes are perpetuated, leading to discrimination and negative prejudices against certain groups. Hence, this study asserts that all subtle forms of discrimination, including seemingly harmless jokes, should be regarded as online hate. This is based on previous research that has highlighted how repeated exposure to such jokes can reinforce racist attitudes [49–51] and despite their seemingly innocuous nature, they can have adverse psychological effects on some individuals [52].

2.2.3 Rules for Classifying Online Hate

In order to gain further insight into the identification of online hate, a list of classification rules described in previous academic studies for hate detection has been compiled below. Considering these rules, a particular post contains hateful content when:

- It highlights a person belonging to a specific group and references a commonly held negative stereotype associated with that group. [41].
- It makes generalised negative statements about minority groups due to the incitement of a negative bias toward the group, for instance, “the refugees will live off our money”. However, there were some authors [35] who expressed uncertainty about whether this particular example constituted hate speech.
- It uses derogatory language and racial slurs with the intention of causing harm or offense [23].
- It uses sexist or racial slurs [35].

- It endorses organizations that promote some form of online hate. Although this may not be an explicit verbal attack on another group, it still qualifies as online hate and should be identified as such. For instance, endorsing an extremist organisation would still be considered as online hate and should be removed or reviewed further. In this aspect, this opposes the perspective of some other authors [41].
- It displays discrimination towards certain countries or religions, though speaking badly about them in general is allowed.[23]
- It asserts the superiority of the in-group through statements. [41].

In addition to this, it is important to distinguish the circumstances where online hate is not present. For instance, this is the case when certain offensive terms are being discussed, e.g. the meaning or origin of a particular derogatory term is being explained, as noted by Warner and Hirchberg in [41]. Furthermore, referring to a particular hate organisation in general is allowed; several historical articles or other forms of legitimate communication may mention organisations such as the Ku Klux Klan or ISIS, though these would not be considered as online hate. It is also worth noting that words such as 'black', 'white', or 'filthy' do not bear any racial undertones on their own, and thus can only be used to identify hateful content in certain contexts and circumstances [53].

2.2.4 Other Concepts Related to Online Hate

When examining the concept of online hate, it is also important to compare it to other related concepts, and understand the various applications of hateful content. Whilst reviewing existing literature within this field, several additional concepts are also introduced or used to exemplify instances of online hate. Some of these related concepts include cyber-bullying, discrimination, flaming, extremism, and radicalisation. A definition for each of these concepts and their relation to online hate are provided in Table 2.2, along with previous studies that have analysed hate in the context of these specific concepts.

Table 2.2: A description of concepts related to online hate and previous studies that have analysed hate within these contexts.

Concept & Example Studies	**Definition**	**Relation to Online Hate**
Cyber-bullying [7, 54]	Aggressive and intentional acts carried out by a group or an individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend themselves [55].	Online hate is not necessarily targeted at a specific person and can be more generic in conduct.
Discrimination [56–58]	A process through which a difference is identified and then used as the basis of unfair treatment [59].	Online hate is a form of discrimination, through verbal means.
Flaming [60, 61]	The outburst of rude or emotional behaviours by specific users. This includes negative behaviours ranging from becoming angry to displaying strong emotions such as hostility [62]. In other words, flaming encompasses a wide range of online behaviours such as abuse, the use of coarse language, slander, the spread of false information, personal attacks, the plastering of message boards with malicious comments and sexual harassment [63].	Hate speech can manifest in any setting, whereas flaming is directed towards a participant within the particular context of a discussion [23].
Online Extremism [26, 64, 65]	The promotion of ideology associated with extremists or hate groups, endorsing violence, often aiming to segment populations and reclaiming status, where out-groups are presented both as perpetrators or inferior populations [66].	Extremist discourses frequently use online hate as a strategy for promoting their content. However, these discourses oftentimes focus on aspects other than perpetuating hate, including the recruitment of new members, government and social media demonization of the in-group and persuasion [67].
Online Radicalisation [8, 68]	Online radicalisation is very similar to the concept of online extremism and is essentially a process whereby individuals through their online interactions and exposures to various types of internet context, come to view violence as a legitimate method of solving social and political conflicts [16, 69]	Radical discourses, including extremism, can employ hate speech. However, topics such as war, religion, and negative emotions are frequently discussed in radical discourses, whereas hate speech may be more nuanced and based on stereotypes [16].

Studying such concepts often provides relevant and useful use-cases to gain insight into methods for detecting and analysing hateful content, and also to analyse behaviours of organised hate. Some of these concepts will, therefore, be used within this research. In particular, the methodology used in this thesis will involve analysing the online behaviours of various hate groups that have been identified as extremist or discriminatory, and will be predominantly applied to hateful content within the context of these two concepts.

2.3 A Review of Existing Research on Online Hate

As previously mentioned, several studies have explored the subject of online hate and suggested definitions and methods for its detection and analysis. This research has been carried out using various qualitative and quantitative approaches from CSS. The remainder of this chapter will discuss and detail some of the findings and limitations of such research, and the different approaches used to provide insight or possible solutions for the phenomenon of online hate.

As this project will largely make use of computational methods to address the research questions described in Chapter 1, there will be particular emphasis and detail on literature using quantitative methods. However, since many of these approaches rely on qualitative methods to validate their findings, it is still important to survey such work to gain a more comprehensive understanding of online hate and its impacts, as well as how any proposed analysis frameworks can work to mitigate this.

2.3.1 Methods for the Detection and Analysis of Online Hate

Conceptual Frameworks of Online Hate

To further build upon the definitions of online hate and related concepts presented above, several academics from CSS-related fields have emphasised on producing conceptual frameworks to aid in understanding its impacts; more specifically, how it can lead to a normalisation of biased behaviours that could potentially lead to violent

hateful crimes. Research carried out by the Anti-Defamation League[1], depicts this process of normalisation through a "Pyramid of Hate", as shown in Figure 2.1, where biased behaviours grow in complexity from the bottom of the pyramid to the top [70].

Although the behaviors at each level negatively impact individuals and groups, as one moves up the pyramid, the behaviors have more life-threatening consequences. Presenting this as a pyramid implies the notion that, as with physical pyramids, the upper levels are supported by lower levels. If people or institutions treat behaviors on these lower levels as being acceptable or "normal", it results in the behaviors at the next level becoming more accepted, and thus even benign instances of online hate could potentially lead to more violent acts of hate offline as extreme as genocide.

Figure 2.1: A "Pyramid of Hate", as proposed in [70], depicting how the normalisation of biased, hateful behaviours can lead to violent hateful crimes.

This is further echoed by Jubany and Roiha when they argue that online hate contributes to stigmatising, marginalising, and intimidating members of distinct and vulnerable groups, whilst simultaneously aiming to reinforce a sense of a group

[1]https://www.adl.org/

under threat from "the others" by uniting like-minded persons [71]. Online hate differs from offline hate in several ways including the length of time it can remain in the public domain online, the perceived sense of anonymity that can make malicious actors more comfortable to spread hateful content, as well as the transnational reach of online environments that can increase its impacts. Regardless of this, it would be false to adopt the assumption that anything happening online is separate from the offline, also known as "digital dualism", as argued by Jurgensen [72]. Though the Internet may expedite and ease interactions for hateful groups and individuals, its role should not necessarily be viewed as an alternative to physical, offline interactions. Moving away from a digital duelist perspective also means putting an emphasis on contextualising online hate, where hate on the Internet could be an expression of a larger phenomenon of increasing hate in society [71].

A further aspect in such research which is often explored is the legal impacts and responses to online hate. When approaching online hate through a legal-normative framework centred around a discourse of violent hate crimes, two problems are highlighted. Firstly, the assortment of new forms of online activity that has emerged in recent years defies the easy categorisation of speech and behaviours that are acceptable and that are not. In other words, unlike activities such as violent extremism online, new movements like social media hate speech cannot be as easily typecast into binary divisions between legitimate or illegitimate forms of political speech and behaviours [24]. Secondly, researchers working on violent online extremism and hate have often presupposed a universal normative framework, which is not as easily transferable to other examples of online hate across a diversity of global contexts. Consequently, there is a growing need to better understand the multiplicity of situated speech acts and cultures of communication underlying violent online extremism in countries with often radically different socio-political contexts [73].

Building on to this, Ramati further argues that, whilst studying the legal aspects of online hate, it is important to consider the the impacts on the 'freedom of speech', as this an argument often used in defense of online hate [25]. He reasons

that a continuous discourse on the justifications for the rights to privacy and free speech in light of technological and political changes is essential, not only for the advancement of knowledge but also for confronting the challenges to these rights. On the one hand, it is important in order to have a robust legal argument against countries who put severe restrictions on speech and knowledge online and who constantly monitor private online correspondence, such as China or Iran. On the other hand, it is important in order to define the limits to those rights in light of the use of online environments by hateful organisations and individuals. Furthermore, policy-making in this space is increasingly based on the dramatic combination of public pressure and unproven assumptions shaped by anecdotal evidence rather than rigorous, empirical research [74]. Recent research efforts have sought to address this gap by quantifying the nature of Internet use among hateful actors [75], some of which will be discussed in the next subsections.

Collection of Large-scale Datasets

In order to study hateful content online, it is essential to use data-driven research with large-scale, reliable and relevant web data. Acquiring such data itself would traditionally entail a high degree of engineering or technical skills and computational resources. Web data, particularly data from application programming interfaces (APIs), has been an enormous boon for researchers studying online social-media platforms [76–79]. However, following major scandals around data privacy and ethics, mainstream social-media platforms, such as Facebook and Twitter, changed previously permissive data access provisions of their public APIs [80]. As a consequence, these mainstream platforms have gotten better at moderating content, which is necessary for detecting malicious usage of the platform, though this also has an impact on research efforts as there is a limited access to this data.

This “post-API age” is characterized by the deprecation of data resources used for research [81, 82], increased stratification of data access based on social, technical, and financial capital [83, 84], and greater fear of prosecution around violating terms of service in the course of research [85, 86]. Baumgartner et al. [87] further

explain that research into online hate, or even any area of social-media study, is in fact oftentimes gated by data engineering problems that must be overcome before any analysis can proceed. Additionally, despite widespread agreement that recent increases in online hate and hate crimes are due to a globalised toxic culture online operating largely outside mainstream social-media platforms, much of the research on extremist use of social media still focuses on mainstream sites like Facebook or Twitter [13]. Since access to these rapidly-changing online spaces is difficult, researchers often rely on out-of-date data. Thus resulting in the recognition of datasets of social-media posts as valuable research contributions in and of themselves.

Several studies [13, 87, 88] have included aspects of scraping from various social-media platforms to form large-scale datasets of social-media data for the purpose of research in CSS. The majority of these papers [15, 89, 90] and the datasets included within them are generally conducted over a single online platform, mostly due to the vast amount of time, resources and efforts required in collecting the datasets. However, some of these papers [14, 91] have noted that it would be insightful to use these datasets in conjunction with datasets from other studies — collected from other platforms, over different time frames, or with data in a different language — to understand the wider context of online hate and the activities of hate groups on social media. These sentiments are reiterated by Baumgartner et al. [13] and Burris et al. [92] when they note that one of the most significant challenges faced by researchers of online hate is gaining a comprehensive set of data to get a wider, clearer understanding of this field. Furthermore, as with any type of scientific data, there are potential validity and measurement problems with such social-media datasets that are best solved through corroboration with other data.

Table 2.3 summarises the details and contributions of some of the existing scientific papers, where the main contribution is the dataset provided.

It should be noted here that only those papers which have published their datasets have been included, which have mostly been published from 2018. These datasets are comprehensive collections of social-media data, which include use cases applicable to online-hate research but are intended for use in a variety of different research

Table 2.3: A summary of academic dataset papers and their contributions including a description of the dataset, the year the dataset was published, the size of the dataset, the time period over which the dataset was collected (Time), the platform collected from and the language of the data collected (Platform, Language), and the source of the dataset (Paper).

Description	Year	Size	Time	Platform, Language	Paper
Long running discussions about the Brexit referendum on Twitter.	2020	50.8M tweets, 3.97M users	Jan 2016 – Sep 2019	Twitter, English	[93]
Thematically distinct datasets of general threads on the /pol/ forum on 4chan.	2020	58K opening posts, 13.6M replies, divided over 329 distinct themes	Dec 2013 – May 2019	4chan, English	[88]
A dataset of Telegram messages collected from publicly-viewable channels	2020	317M messages from 2.2M users across 28K channels	Oct 2015 – Oct 2019	Telegram, English	[87]
A dataset (continuously updated in real-time) of posts from Reddit retrieved using the Pushshift, a social media collection, analysis and archiving platform.	2020	651.7M submissions, 5.6B comments posted on 2.8M subreddits	June 2005 – April 2019	Reddit, English	[13]
A dataset of threads and posts retrieved from the /pol/ forum from 4chan.	2020	134.5M posts, 3.3M conversation threads	June 2016 – Nov 2019	4chan, English	[14]
A dataset of posts, comments and user profiles scraped from Gab.	2019	37M posts, 24.5M comments, 820K user profiles	Aug 2016 – Dec 2018	Gab, English	[89]
A data collection pipeline and a dataset of news articles and their associated sharing activity.	2019	331K articles, 37K users, 975K tweets	Jan 2018 – Sep 2018	Twitter, English	[94]
A dataset of posts from Mastodon, including those labelled inappropriate by the users who posted them.	2019	5M posts	March 2016 – Jan 2019	Mastodon, English	[15]
A dataset of posts and user profiles collected from Gab.	2018	22M posts, 336K users	Aug 2016 – Jan 2018	Gab, English	[91]
An annotated dataset of abusive tweets.	2018	80K tweets	March 2017 – April 2017	Twitter, English	[90]
An anonymised dataset of messages from public groups on WhatsApp.	2018	454K messages, 45.7K users, 178 public groups	May 2017 – Oct 2017	WhatsApp, English	[95]

fields, hence many of these dataset papers are not hate focussed. It is also worth noting that, in addition to those published through scientific papers, public datasets of social-media data are sometimes published on various data science communities and crowd-sourcing sites as well. For instance, the data science community Kaggle[2] has previously published a collection of over 17K pro-ISIS tweets retrieved from 112 distinct pro-ISIS Twitter accounts. Such datasets were not included here as this subsection primarily focuses on how collecting and publishing social-media datasets has become an academic contribution itself, however, some of these are used and acknowledged throughout the rest of this thesis.

Automatic Detection of Hate Speech

One of the most common aims of using computational approaches to study online hate is to be able to detect hateful content, and then consequently remove it, in online environments automatically and accurately. Automatic content removal has been highly successful in detecting and removing certain types of content. Facebook reports that it automatically removes 99% of ISIS and al-Qaeda material uploaded to the platform [96]. Similarly, YouTube reports that 98% of the videos that are removed for violating violent-extremism rules are detected by an automated system [97]. During the first six months of 2019, Twitter removed 584,429 accounts for violations related to hateful conduct, and 87% of these accounts were flagged for removal by automated tools [98]. This undoubtedly represents a success for major social-media platforms [99]. Automatic hate-detection tools are allowing more content to be removed, and for this to be done more quickly and with greater coverage.

For this reason, several academic studies within this field make use of machine-learning algorithms including Support Vector Machines (SVM) [42, 56, 100–102], Random Forests [16, 103–106], decision trees [16–18, 107, 108], logistic regression [19, 109–112], Naive Bayes [53, 113–116] and, more recently, deep learning [105, 117–120] with varying levels of accuracy. It should be noted here that it is often difficult

[2]https://www.kaggle.com/

to compare the performance of different machine learning models across various studies due to different datasets being used, different platforms being analysed, and different features being chosen to classify hate speech. Various metrics are also used to estimate the performance of the models including Precision, Recall and F-scores[3] — which are the most commonly used metrics — as well as Accuracy[4] and Area Under the Curve (AUC)[5].

In addition to the application of different algorithms, most of the studies in this literature review also use different types or combinations of Natural Language Processing (NLP) techniques within their text mining approaches. Finding the most appropriate features to train the classifier is often the aspect which has the most impact on the accuracy of a particular model in machine learning. Further detail of some of the typical types of features obtained through popular NLP techniques are provided below. It is worth noting that, while automatic detection of online hate is an ever-increasing field of study with promising results, it is not always as clear-cut as detecting and removing hateful content, especially in the case of high-profile social-media accounts. These sentiments are emphasised by Houseley et al. [1], where they note that such high-profile posts can be identified as 'ignition points' for the generation of antagonistic communication flows. Consequently, the removal process can take longer in these cases.

The majority of the articles reviewed attempt to adapt strategies already known in text mining to the automatic detection of hate speech. Table 2.4 provides a summary of some of these studies and the algorithms and features that were used. A description of some of the commonly used algorithms is also provided below.

[3]**Precision** refers to the fraction of positive predictions that are correct among the total positive predictions; **Recall** refers to the percentage of positive predictions among the total number of positive instances in the dataset; and the **F-score** is the balanced average of the precision and recall measurements.

[4]**Accuracy** is simply the ration of correct predictions (both positive and negative) to the total number of predictions.

[5]The **AUC** provides a means of calculating the rank correlation between predictions and targets, and is used to show how good the machine learning model is at ranking predictions

Table 2.4: A summary of academic studies that make use of automatic hate speech detection, which includes the year of the study (Year), the algorithms used (Algorithm), the types or combinations of features used (Features), the number of posts analysed (Posts), the platform(s) analysed (Plat.), the language of the data analysed (Lang.), the metrics used to estimate the performance (Perf.), and the reference to the study (Paper).

Year	Algorithm	Features	Posts	Plat.	Lang.	Perf.	Paper
2020	Deep Learning (LSTM)	Word embeddings (word2vec)	25,000 tweets	Twitter	English	Precision=0.920, Recall=0.920, F-score=0.915	[107]
2020	SVM	Character n-grams	15,702 tweets	Twitter	English	Accuracy=0.646	[110]
2020	Deep Learning (RNN, GRU)	Word embeddings (word2vec)	14,266 posts	Facebook	Amharic	Accuracy=0.9256, AUC=0.9785	[119]
2020	Deep Learning (LTSM and Char-CNN)	Word embeddings (word2vec)	3 sets: 15476, 24783, 15001 tweets	Twitter	English	Accuracy = 91.09, 84.14, 58.83 resp.	[120]
2019	SVM	Lexical syntactic, Character n-grams	3 sets: 3251, 4000, 5006 tweets	Twitter	English	Accuracy= 0.7605, 0.7947, 0.8937 resp.	[56]
2019	Random Forest	Rule-based approach	14,509 tweets	Twitter	English	Precision=0.711, Recall=0.712, F-score=0.713	[103]
2019	Naive Bayes	TF-IDF, word embeddings (count2vec)	12,805 tweets(Eng), 5384 tweets(Span)	Twitter	English, Spanish	F-score= 0.76(English), 0.77(Spanish)	[108]
2019	Naive Bayes	Word embeddings (count2vec), TF-IDF	1339 posts	Facebook	Bangla	Precision=0.75, Recall=0.71, F-score=0.73	[115]
2017	SVM, Logistic Regression	TF-IDF, POS	24,902 tweets	Twitter	English	Precision=0.91, Recall=0.90, F-score=0.90	[42]
2017	SVM, deep learning (LSTM)	Sentiment, Word embeddings (word2vec)	3575 posts	Facebook	Italian	Precision=0.833, Recall=0.872, F-score=0.851	[102]
2016	Logistic Regression	Character n-grams	16,914 tweets	Twitter	English	Precision=0.72, Recall=0.77, F-score=0.73	[112]
2016	SVM, Random Forest	BOW, Typed Dependencies	5593 tweets	Twitter	English	Precision=0.79, Recall=0.59, F-score=0.68	[106]
2015	Logistic Regression	Paragraph embeddings (paragraph2vec)	266,056 posts	Yahoo	English	AUC = 0.80	[111]

Dictionaries. Dictionaries are often utilised in text mining through the creation of a list of words (i.e., the dictionary) to be counted within some text. These frequencies can then be used as features or scores, with an additional normalisation step by considering the total number of words in each comment. Regular expressions can also be used with this approach. In the case of hate speech detection, this approach has been conducted in previous studies using content words [121], the number of profane words [122], frequently used forms of verbal abuse and stereotypical utterances [123], and Ortony Lexicon [123].

Bag-of-words (BOW). A similar model to dictionaries is the bag-of-words approach [17, 53, 124]. Instead of using a predefined set of words, a corpus is created based on the words in the training data. The frequency of each word is then used as a feature for training a classifier. However, a disadvantage of this approach is that it ignores the word sequence and may lose syntactic and semantic content, potentially leading to misclassification. To address this limitation, N-grams can be used.

N-grams. N-grams are commonly used in automatic hate-speech detection and related tasks, and have been applied in various studies [40, 42, 105, 106, 112, 116, 124]. This approach involves combining sequential words into lists of size N, where all expressions of size N are enumerated and counted to improve classifier performance by incorporating context. N-grams can also be used with characters or syllables, which is less affected by spelling variations than using words. These approaches have often proven to be more predictive than token N-gram features, for the specific problem of abusive language detection [125]. However, using N-grams has its drawbacks. Related words may have greater distance between them in a given sentence [106], and increasing the N value to solve this issue can slow down processing speed [55]. Studies also suggest that higher N values (around 5) perform better than lower values (such as unigrams and trigrams) [116], and combining N-grams with other features can further improve performance [126].

TF-IDF. The Term Frequency-Inverse Document Frequency (TF-IDF) approach has also been utilised for classification purposes [123]. TF-IDF measures the significance of a word in a document within a corpus, and is directly proportional

to the number of times that a word appears in the document. However, unlike the bag of words or N-grams, TF-IDF adjusts the frequency of the term by considering the frequency of the word in the corpus, thereby offsetting the fact that certain words are more commonly used in general, such as stop words.

Topic Classification. Topic Classification or Topic Modelling involves using these features described above to identify the underlying topics present in a given document or text. An example of this can be seen in the work of Agarwal and Sureka [16], where linguistic features for topic modeling were employed to detect posts related to specific topics, including Race and Religion.

Sentiment. Considering that hate speech generally has a negative tone, researchers also use sentiment analysis as a feature to identify hate speech in text [16, 42, 102, 116, 121, 127]. Though such studies have shown promising results when identifying the polarity of content, autonomous sentiment analysis methods are still limited when it comes to differentiating such content from sarcasm. Within the literature reviewed, this type of feature is usually used in combination with others that proved to improve accuracy results of the classifier [116].

Word Embeddings. This is another popular approach of representing the vocabulary in a particular document or piece of text. Here, word embeddings refer to a class of techniques where individual words are represented as vectors in a predefined vector space; one such technique is word2vec [128]. The objective is to have words with similar context occupy close spacial positions. However, hate speech detection involves the classification of sentences or passages rather than individual words. One possible solution for this is to average the vectors of all words occurring in a passage or sentence [126], though this method has limited effectiveness [40]. Alternatively, Djuric et al. [111] suggest a paragraph2vec technique for categorizing user comments as either 'clean' or 'abusive' and to predict the central word in the message. This method of *paragraph embeddings* has demonstrated greater efficacy [40].

In addition to the approaches described above, other features have also been used to classify online hate speech. Such approaches were based in techniques such as Named Entity Recognition (NER) [129], Word Sense Disambiguation Techniques

to check Polarity [40, 127], frequencies of personal pronouns in the first and second person [130], the presence of emoticons [122, 131], as well as capital letters [132]. Other characteristics of the data being analysed were also considered such as hashtags, mentions, retweets, URLs, number of tags, terms used in the tags, number of notes (reblog and like count), and link to multimedia content, such as images, videos, or audio attached to the post [16, 23] — these characteristics vary depending on the platform and type of dataset being analysed.

Content Diffusion Analysis

Another computational analysis approach that is being increasingly used in recent years is the analysis of the diffusion dynamics of online hate. The concept of diffusion dynamics has been studied and discussed for several decades to investigate the flow of online content through several different sources, and how the audiences then interpret this content. Thus there is extensive literature in CSS exploring the diffusion of cultural fads [133], innovation [134], or products [135]. From this, several theories, frameworks and models have been developed on the process of information diffusion applied across various disciplines.

In his theory "Diffusion of Innovations", Rogers defines diffusion as the process by which an innovation is communicated over time among the individuals in a social system, where the innovation must be widely adopted in order to self-sustain. This theory proposes that diffusion consists of four main elements influencing the spread of a new concept or idea: the innovation itself, the communication channels used, the amount of time taken for the concept to be adopted, and the members of a social system [134]. Through viewing the diffusion process as the transmission of ideas between individuals, various contagion models have also been adapted from mathematical epidemiology, some of which are described below [136].

One of the most widely adapted models is the Bass model [137], which has become an important exemplar in marketing science. The Bass model is an independent interaction model that formalises the insight that innovation diffusion is a two-step flow process, where media influences "innovative opinion leaders" to adopt some

new product or idea, who in turn influence people imitating their behaviour to adopt the product or idea as well [138, 139]. This model has been expanded in several following studies to create threshold models of diffusion [140, 141], where the idea is that the initial number of innovators adopting the product or concept needs to be above a certain threshold in order for the innovation diffusion to be successful. Dodds and Watts propose a further model which incorporates both the above types of contagion models. This introduces memory of past exposures to a contagious influence to generalise and interpolate between independent interaction, as depicted in the Bass model, and threshold models of contagion [142].

In addition to modelling the statistical properties of diffusion, social perspectives and implications have also been explored. For example, Bikhchandani et al. studied individuals' tendency to learn from and imitate others and explained the role of this behaviour in the diffusion of cultural fads [133]. Theories such as the critical mass theory proposed by Markus [143] provide a more holistic insight into the interactive diffusion of new concepts and products within communities, and describe the relationship between the growth of adoptions and network externality. Other studies provide a more grass-root understanding of diffusion within various use cases and communities by modeling adoption and coordination processes with various influencing factors, including the structures of the networks and communities being investigated [144–146].

The ever-increasing usage of online social-networks have provided a vast amount of user behaviour and interactions for observation, and have thus enabled researchers to conduct more large-scale studies of diffusion patterns and dynamics. For instance, Buhl, Günther and Quandt [147] observe and analyse the diffusion dynamics of the online news ecosystem, with the aim of describing generalised patterns of news diffusion processes among online news-sites in terms of their range, duration, and dynamics. Another example of such a study was conducted by Leskovec et al. [135], which discerned a power-law distribution of online recommendation cascades and proposed a model to identify community structure, product features, and pricing strategies for successful virtual marketing.

Within online social-networks, the process of diffusion often involves behaviours that seem more menial, and thus even the act of joining a particular group can contribute to the diffusion process. Within their study of friendship links and community membership on LiveJournal, and co-authorship and conference publications in DBLP, Backstrom et al. found that whether an individual will join a group depends not only on how many friends one has within the group, but also how these friends are connected to one another [148]. Further to this, Kim, Baek and Kim [149] analyse frames of news along with the formation of opinions and how they diffuse online, where they found that public discourse on the Internet is clearly "polarized" and "fragmented" along political ideological lines.

More recent studies have investigated the diffusion of fake news [32, 33, 150], re-tweet cascades [32, 151–153], and rumours [154–158]. Cheng et al. [152] perform a large scale-analysis of recurring cascades in Facebook, where they observe that content virality is the main driver for recurrence. In a similar study conducted by Del Vicario et al. [154], the authors investigated how Facebook users consumed information related to scientific discourse and conspiracy theories. Here, they observe that selective exposure to content is the primary driver of content diffusion and generates the formation of homogeneous clusters, or echo chambers, which encourage and enforce imitation.

Furthermore, other research has attempted to measure the virality of a cascade [159] and whether content would disseminate in specific demographic groups [160]. For instance, previous studies have found that radicals are far more likely to post political content than moderate political users and that users with more followers are less likely to post racist content [53, 161]. Romero, Meeder, and Kleinberg provide a comprehensive examination of how diffusion cascades on Twitter differed for various topics like politics, sports, movies, etc. [162]. Their research found significant differences in the distribution of cascade sizes and persistence across topic areas, arguing that repeated exposure to specific types of content made users more likely to share it. A following study by Wu also demonstrated that

information cascades carrying negative sentiments fade far more rapidly than their positive sentiment counterparts [163].

Similarly, using a large sample of social-media interactions concerning polarising issues in public policy debates (including gun control, same-sex marriage and climate change), Brady et al. find that the presence of moral-emotional language in political messages substantially increased their diffusion within ideological group boundaries [34]. These findings offer insights into how moral ideas spread within networks during real political discussion. The recent research of Vosoughi, Roy and Aral, provides a similar examination of the spread of true and false rumours. The authors analyse the diffusion cascades of verified true and false news stories on Twitter from 2006 to 2017 over 125,000 total cascades [32]. The study finds major differences between the cascades of true and false stories: false rumours spread faster and more broadly than their true counterparts. Interestingly, the authors demonstrate that this difference in sharing behaviour occurs in human Twitter users, not bots on the platform, and that motivation to share novel news stories drives a lot of the virality of this fake information.

Although this concept of diffusion has been explored substantially in online communities, particularly those on social media, there are few studies that apply this to the phenomenon of online hate. One such study is conducted by Beatty, who proposes a framework for hate speech diffusion cascades on Twitter using the network topology and time dynamics of the spread of the content [20]. This framework was developed to compliment previous text-based classifiers by providing the sense of the context of a message by examining how other users interact with it, though it does not enhance the detection of hate speech. More recently, Mathew et al. provide further insight by exploring the diffusion dynamics of hateful and non-hateful users in 21 million posts collected from the online social-media platform Gab [21]. Their work finds that the content generated by the hateful users tend to spread faster, farther and reach a much wider audience as compared to the content generated by normal users. They also find that hateful users are far more densely connected among themselves.

Network Analysis

Network analysis is an approach used for online hate analysis in oftentimes similar ways to the approach of content diffusion analysis. However, the difference between these two approaches lies in *how* they are used. In this case, network-analysis techniques are used to detect and remove online hate by identifying networks of hateful communities and disrupting these networks, for instance by blocking the most influential nodes or clusters. In contrast, content diffusion, as explained above, describes the process and path through which content is disseminated and propagated online, thus disrupting this content flow is another method that could be used to counter online hate. This section will be providing some detail on approaches used to detect, disrupt or remove networks of hate by discussing previous literature using this approach.

As previously noted, the Internet has become a central aspect of daily life and, as a result, interaction is increasingly mediated by the online medium [164]. However, Keipi et al. further argue that while these online environments have provided great affordances such as exposure to a greater wealth of information, a perceived sense of anonymity, and much larger and more dynamic social networks, these same affordances ironically play a highly influential role in dealing with online hate [164]. Developing a more complete view of dangerous pockets in social networks and how they are reinforced by the pattern recognition of the online environment could lend itself to more effective methods of meeting identity and expressional needs, without nurturing cycles of hate and social harm. Network-analysis approaches would therefore be essential to understand the structures and dynamics of these networks. This generally involves the mapping of relations that connect online communities together, consequently drawing out components and associations between individuals. The resulting network is then formed of 'nodes' (actors, organisations, or events) and the links between them.

Figure 2.2, mainly adapted from the observations made by Durland and Fredericks [165], provides more detail on the variables measured during network analysis to provide this insight.

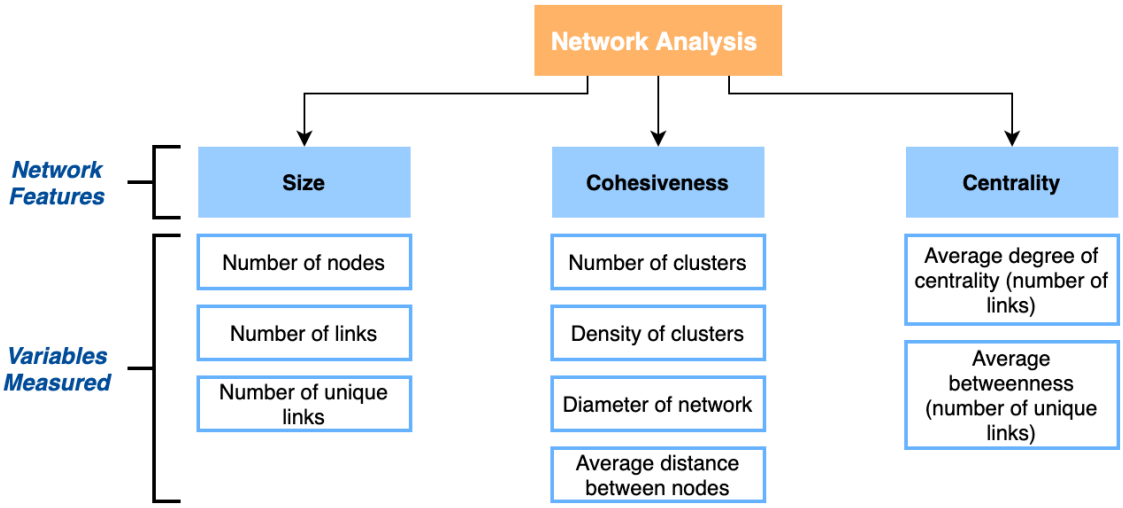

Figure 2.2: The various features of networks that are measured in network-analysis approaches.

Snow, Zürcher, and Ekland-Olson first proposed that rigorous analysis of the structural properties of social networks could be valuable for the study of social movements in 1980 [166]. Following on from this work, several researchers have sought to apply network concepts to a variety of social movements, and network characteristics have come to occupy a larger place in theories of collective behaviour [167–174]. One such study conducted by Rosenthall et al. [167] examines organisational affiliations of 19th-century women reform leaders in New York State as a case study of networks among social movements. Here the authors used network-analysis approaches to map the interconnections between organisations, measure the intensity of these interconnections, highlight clusters of proximate organizations, and identify the organisations central to the clusters. Kim and Bearman further developed a dynamic network model of collective action, where they found that network density played a crucial role in shaping the outcomes of social movements [174].

Though the above listed literature made important contributions to the understanding of social movements, it remains limited in two respects. First, most of the applications of network analysis to social-movement research focus on interpersonal ties. Less attention has been given to the structural properties of networks formed through ties among collective groups, organisations, or other corporate entities. Second, network characteristics are usually measured at the individual level in

terms of the number and type of ties maintained by individual actors. Although investigating the question of which individuals would join these social movements led to an understandable focus on these individuals as the only units of analysis, previous theoretical literature also notes the importance of inter-organisational ties for the dynamics of social movements [175–177]. Methods of social-network analysis that allow one to map the structural properties of social networks in their entirety is therefore needed.

These limitations are pointed out and addressed by Burris, Smith and Strahm in [92], which is also one of the pioneer studies applying network-analysis approaches to the study of online hate. Here, the authors use methods of social-network analysis to examine the inter-organisational structure of the white-supremacist movement. This study departs from most previous applications of network concepts to social-movement research by examining ties among movement leaders and organizations rather than purely interpersonal ties, where links between websites are treated as ties of affinity, communication, or potential coordination. Through this, the authors observe the apparent decentralised structure of white supremacist movements with multiple influential centres, as well as how the website links connect various groups in many countries, showing that these networks transcend regional and national boundaries. This methodology is also used by Zhou et al. [178] to understand the dynamic of US domestic extremist and hate groups, and by Tateo [179] to explore the structure of the Italian far-right network online.

As the Internet became an increasingly more interactive environment, the development of online social communities became more present in society through blogs and social-media platforms. While network-analysis techniques were previously used in several studies, as listed above, to analyse the structure of web sites of hate groups on the Internet, Chau and Xu [4] point out that the ever-growing use of blogs to express opinions has led to the emergence of many more online communities, including those of hate groups. To address this, the authors applied network-analysis approaches to investigate the structural properties of the social networks of bloggers in these hate groups. A similar methodology is used by

Bermingham et al. to examine social relationships and networks in the context of online radicalisation by jihadi extremists within YouTube profiles and comments [69]. Like previous research, both these studies find that the structures of hateful networks are largely decentralised, though Bermingham et al. do point out that networks of female users have a higher density than those of male users, indicating a potentially increased leadership role for women online than they would generally be held to have within jihadi circles [69].

More recently, researchers have explored a variety of different platforms enabling the formation of online hate or extremist networks, and thus drawn various conclusions when studying the environments of these online communities. In addition to presenting the structural aspects of their communities, network analysis approaches have provided insight into the strategies and tactics used by hateful groups or individuals to effectively promote their ideas through the networks, as well as to recruit new followers. For instance, Schwemmer makes use of network-analysis techniques to monitor the German right-wing movement Pegida on Facebook for 18 months [22]. Through this study, the author finds that, over the analysis time-frame, the movement created more radical and xenophobic content, which in turn attracted more followers. In contrast, an article written by Weaver argues that such right-wing organisations are increasingly revising their strategies and editing the content they post online, so as to keep it consistent with restrictions put in place by online-platform providers, hence preventing their content from being removed [180].

Additionally, as well as analysing prominent accounts of hateful organisations or individuals, a number of studies have argued that the various functionalities of social-media platforms also facilitate the development of hateful communities, and thus should also be explored to analyse networks of hate. One such study that used this methodology is conducted by Eddington, where the "Make America Great Again" campaign slogan, first used by Donald Trump in the 2016 US presidential elections, is investigated on Twitter via the #MAGA hashtag [3]. Though this hashtag was initially introduced to allow social-media users to connect with his campaign, the presence of hate groups connecting with the hashtag also

became increasingly apparent. Exploring the networks of this hashtag showed how it was used as a communicative organising site for white-supremacist groups and illuminated the overtly far-right content and hashtag conversation spaces shared and embedded within #MakeAmericaGreatAgain [3]. This demonstrates the importance of establishing the various ways in which networks of hateful content can be formed and organised on social media to gain a more complete idea of the structures of these communities.

One of the most comprehensive studies to explore the networks of online hate groups was conducted by Johnson et al. [11]. In this study, the authors examine the dynamics of hateful communities on both Facebook and VKontakte over a period of a few months. Through this approach, the authors develop a mathematical model showing that online hate groups are organised in a "network of networks" structure of highly resilient clusters, which are globally interconnected via "hate highways" facilitating the spread of online hate across different countries, continents and languages. Using their model, the authors additionally demonstrate that banning hate content on a single platform could aggravate online hate ecosystems and promote the creation of clusters that are not detectable by platform policing, leaving hate to thrive unchecked.

Kashpur et al. [28] brought further insight into the networks of online hate by investigating where network leaders and frequently present members of Russian nationalist online groups go once these groups have been blocked or removed, and which thematic communities they use to continue their activity. Their findings demonstrate that the blocking of larger far-right groups leads to an increase in the number of smaller shelter groups that maintain their ideological basis, though use much less obvious hate speech, thus affirming obeservations made by Johnson et al. [11].

2.3.2 Cross-platform Analysis of Online Hate

Despite all the extensive research and methods used to analyse online hate described above, very few studies have investigated how hateful behaviours and content

compare and relate across different online platforms [181]. It has only recently been recognised within academic literature that online hate does not simply exist independently on online platforms, rather networks of hate are often linked across these platforms, forming a global 'network of networks' dynamic [11].

In other words, online hate thrives globally through self-organized, scalable clusters that interconnect to form networks spread across multiple social-media platforms, countries and languages. These networks formed by hate groups have proven to be remarkably resilient, and have increasingly shown to migrate across various platforms and other networks, maintaining and oftentimes expanding their connections in the process. This interconnection of several hate clusters allows for the rapid rewiring and self-repairing of the network at the micro-level when it is attacked [11]. Though there is sufficient evidence that suggests and proves such strategic usage of multiple platforms by hate groups, minimal research has been carried out to explore this further.

One of the first studies to explore the wider network of online platforms, and how various web communities impact and influence each other was carried out by Zannettou et al.; in this study, the authors study how mainstream and alternative news propagate across multiple online communities, whilst measuring the influence that each community has on each other [29]. Using a statistical model, they highlight that small "fringe" online communities within Reddit and 4chan can have a substantial impact on large mainstream online communities like Twitter. The authors demonstrate how such online platforms are clearly not independent; while they do exhibit different behaviours and internal influence, they are also affected by each other, as well as by the greater Web [29]. Further insight into cross-platform behaviours is provided by Hine et al. [31], who study hate speech on 4chan's Politically Incorrect board (/pol/). The authors discuss how they found evidence of organised campaigns, called 'raids', that aim to disrupt the regular operation of other online communities; for instance, they show how 4chan users raid YouTube videos by posting large numbers of abusive comments in a relatively small period of time.

In the context of automatic hate detection online, there is a lack of development and testing of models using data from multiple social-media platforms. Instead, studies tend to focus on one platform. This mono-platform focus is problematic because there are no guarantees the models that researchers develop generalise well across platforms. It is, therefore, a reasonable assumption that developing a universal hate classifier could benefit from the information retrieved from various training sets and contexts. Furthermore, the lack of universal classifiers means that the results across studies and social-media platforms are not easily comparable. Out of the existing studies that do use data from more than one platform, Silva et al. [182] and Mondal et al. [44] evaluate their results for automatic hate speech detection on datasets from two platforms (Twitter and Whisper), achieving reasonable performance on both datasets.

In addition to this, certain studies make use of multiple platforms to test the performance of detection models [44, 182], though few studies make use of multiple platforms for both the model development and evaluation. One of the few studies to do this was conducted by Chandrasekharan et al. [183], who propose an approach, named"Bag of Communities", that uses training data from nine communities within 4chan, Reddit, Voat, and Metafilter, to identify abusive content within online communities. Through applying this to common problems faced by several platforms and content moderators, their approach out-performs others that only concentrate on one community. A further, more recent study to explore automatic hate detection across multiple platforms is conducted by Salminen et al. [43]. In this study, the authors create a hate speech detection model with datasets from Twitter, YouTube, Wikipedia and Reddit, and find that generalising the model to multiple social-media platforms is good, though this varies slightly between the platforms.

Although, research within this aspect of online hate is scarce, in the last year, some studies have realised the importance of the insights that can be gained from cross-platform analysis. In addition to universal frameworks for automatic hate detection, the presence of hate across multiple social-media platforms poses another important question within this field of research: do hate groups use

different social-media platforms in different ways? However, in order to answer this question effectively, it it essential to have comprehensive data from various online environments for analysis. With this motivation, Phadke and Chandaluri conduct a preliminary study where they identify various online hate groups and their Twitter, Facebook, and YouTube accounts, and collect data over the same time period across these platforms [27]. They also present selected linguistic, social engagement, and informational trends, including how the number of certain types of abusive posts increase on different platforms during social movements or around the times of pivotal historic events. For instance, the authors find that the number of anti-Muslim posts around the time of the 9/11 remembrance day increases significantly.

In addition to this, little is known about the communication practices of hate groups across multiple online platforms. Phadke and Mitra utilse the 'framing' theory from social-movement research and analyse domains in shared links to juxtapose the Facebook and Twitter communication of 72 hate groups (designated by the Southern Poverty Law Center) spanning five types of hate ideologies [26]. The findings from this study show that hate groups use Twitter for educating the audience about problems with their targets, maintaining positive self-image by emphasizing their high social-status, and for demanding policy changes to negatively affect their targets. On Facebook, they use fear appeals, call for active participation in group events (membership requests), all while portraying themselves as being oppressed by their targets and failed by the system. This provides some initial insight into how each platform is used differently by hate groups, though further research is required to confirm and support these observations.

2.4 Designing an Analysis Framework

Through examining the vast amount of academic literature in the field of online hate, this chapter has already established that social-media platforms have emerged as a significant source of information and data for researchers across various disciplines seeking to understand the behaviour, attitudes, and opinions of individuals and groups. However, with the immense amount of content available from these resources,

analysing social-media data can often be a challenging and complex task. Therefore, when conducting research on social-media content, it is generally considered a best practice to employ an analysis framework. This helps to ensure that the collected data is comprehensively analysed and correctly interpreted.

A well-designed analysis framework holds significant importance in CSS research in particular due to its ability to provide clear structure and guidance throughout the research process. An analysis framework is a systematic approach that outlines the steps required to analyse data effectively [184, 185] . It plays a vital role in organising the research process, ensuring that the data collected is relevant to the research questions being investigated, and identifying variables and indicators that are important for understanding the phenomenon under study [83]. These factors are integral to enhancing the accuracy and reliability of research findings, particularly when studying complex social phenomena, such as political polarisation [186, 187].

In addition to this, an analysis framework provides researchers with the ability to identify patterns and trends within social-media and online-forum data, allowing for valuable insights into the opinions, attitudes and behaviours of individuals on online platforms [188]. This is particularly important when considering the overwhelming amount of content available on such platforms, which can make it difficult to extract relevant information. This can, in turn, help researchers develop actionable insights that can inform policy decisions, shape interventions, and contribute to the development of a more informed and responsible online community [189].

Social-media data may not always be representative of the general population, and it is essential to identify any limitations in the data to avoid drawing incorrect conclusions. By developing and employing a robust analysis framework, researchers can identify and address any limitations within the data, ensuring that their findings are accurate and unbiased [190]. Developing such a framework also helps to ensure that the analysis process is systematic and transparent, which is imperative for building trust in research findings and ensuring that other researchers can build on existing work [191]. This can further help identify ethical considerations that need to be taken into account when analysing vast amounts of social-media

data, and ensuring any terms-of-service governing the collection and usage of this data are complied to [192]. By following a standardised framework, researchers can maintain the integrity of their studies and ensure that their analyses are consistent and replicable.

Developing a social-media analysis framework requires careful consideration of various criteria and requirements to ensure ensure the effectiveness and validity of the research. Through examining related literature across various disciplines on designing a structured approach to social-media analysis, a list of requirements that should be considered by researchers is presented below:

1. **Clearly defined research objectives:** The research objectives of the analysis framework should be clearly defined and aligned with the goals of the study, including what questions to answer and what insights to provide. This will help in guiding the selection of appropriate data sources, analysis methods, and evaluation criteria.

2. **Selection of relevant data sources:** The data sources used in the analysis framework should be relevant to the research objectives and the target audience or users. For example, if the study is focused on consumer behaviour on social media, platforms such as Twitter and Instagram may be more relevant than LinkedIn [193].

3. **Appropriate data collection methods:** The data collection methods used in the analysis framework should be appropriate to the research objectives and the data sources being analysed; these include web scraping and using the API services provided by the online platforms. The methods should ensure that the data collected is of high quality and is representative of the groups or individuals being studied [194].

4. **Valid and reliable data analysis methods:** The data analysis methods used in the analysis framework should be valid and reliable. This means that they should be appropriate to the research objectives, accurately capture

the data being analysed, and produce reliable results [83]. The selection of analysis methods needs to be determined for various components, which could include the following:

- Data processing: Decide how the data will be processed, such as cleaning and normalisation, language processing, sentiment analysis, and topic modeling.
- Metrics: Determine which metrics will be used to evaluate social media performance, such as engagement rate, reach, and sentiment.
- Visualisations: Determine how the data will be visualised, such as dashboards, reports, or interactive tools.

5. **Validation:** Validate the accuracy of the analysis framework using data samples, manual reviews, or other methods to ensure that the results are reliable and meaningful. The validation criteria used in the framework should be consistent with the research objectives [195].

6. **Ethical considerations:** The development of a social-media analysis framework should take into account ethical considerations such as privacy, informed consent, and data protection laws. Researchers should ensure that they obtain the necessary permissions and adhere to ethical guidelines to protect the privacy and rights of individuals whose data is being analysed [196].

Overall, academic literature suggests that a comprehensive framework is essential for generating accurate, reliable, and meaningful insights from social-media data, and such frameworks should be adaptable, data-driven, and aligned with the objectives of the analysis. The insights gained from these studies were then used to help define the requirements for the analysis framework developed for this thesis, detailed in Chapter 4.

2.5 A Comparison of Existing Online Hate Analysis Frameworks

The literature review presented in this chapter has detailed some of the most common approaches and techniques applied to the study of online hate and social-media content in general. A variety of computational methods have been used to both automatically detect hateful content, including machine learning, as well as perform large-scale analysis, including sentiment analysis and topic modelling. In order to further evaluate how these methods and requirements defined in the previous sections are used in online-hate analysis frameworks, a comparison of existing frameworks presented in academic literature is provided in this section, with the findings from this comparison summarised in Table 2.5.

The frameworks included within this table are those that have been specifically developed for the study of online hate, however several other social-media analysis frameworks and tools also exist within the literature, including the 4CAT [204] and SMAT [205] analysis toolkits. The columns in Table 2.5 outline the criteria used for comparing the frameworks, which were informed by the findings from this literature review. This includes the data sources used by the framework; whether the tool is focused on the detection or analysis of online hate; the methods of analysis that were used; and finally whether any other counter-hate functionalities were used, such as counter-narratives. The insights obtained were then utilized to inform and modify the requirements established in the preceding section for the analysis framework designed in Chapter 4.

Several previous efforts have worked on developing frameworks and tools for law enforcement, platform providers and other researchers to facilitate the detection and analysis of online hate. Many of these frameworks and tools focus on the classification and detection of hateful content. For instance, Davidson et al. [42] developed an online hate-detection tool, HATE Sonar, which is a hate-speech detection library that takes text input and classifies it in one of three categories: hate speech, offensive language, or non-hateful. Another tool developed by researchers at the University of Sheffield, GATE Hate Tagger, provides further classification by tagging any abusive

Table 2.5: A comparison of existing online-hate analysis frameworks presented in literature.

Ref.	Framework Description	Data Source(s)	Detection	Sentiment Analysis	Topic Modelling	SNA	Content Diffusion	Counter-narratives
[42]	HATE Sonar, a hate-speech detection library that classifies hate and offensive language	Twitter	✓					
[197]	GATE Hate Tagger, tags any abusive utterances in text inputs	Twitter	✓					
[198]	A sentiment analysis based web crawler for extremist content	Web pages	✓	✓				
[199]	A hate-detection scheme made up of recurrent neural network classifiers to distinguish racism and sexism from Twitter posts	Twitter	✓					
[200]	ParityBOT, a Twitter bot to counter abusive tweets aimed at women in politics	Twitter	✓					✓
[201]	Hatemeter, a platform that uses NLP, machine learning and big data analytics to identify "red-flags" of anti-Muslim hate speech	Twitter, Facebook	✓		✓			✓
[4]	An approach using semi-automatic techniques to identify and analyse anti-Black hate groups on the Xanga blog hosting site	Xanga blog posts				✓		
[202]	A dynamic network framework to characterise hate communities on Twitter	Twitter	✓			✓		
[21]	A framework that looks into the content diffusion dynamics of posts made by hateful and non-hateful users on Gab	Gab				✓	✓	
[203]	RETINA, a framework used to model and predict the diffusion dynamics of hateful content on Twitter	Twitter			✓		✓	
[29]	A statistical model to study how mainstream and alternative news propagate across multiple online communities	Twitter, Reddit, 4chan				✓	✓	
[183]	An approach to identify abusive content in online communities	4chan, Reddit, Voat, MetaFilter	✓					
[43]	An online hate classifier for multiple social-media platforms	Twitter, YouTube, Reddit, Wikipedia	✓					
[11]	A mathematical model of hateful clusters across Facebook and VKontakte	Facebook, VKontakte				✓		
[16]	A cascaded ensemble learning classifier for identifying Tumblr posts having racist or radicalized intent	Tumblr	✓	✓	✓			

utterances in text inputs. This classification also allows for the differentiation between different types of hate and includes the type of abuse present within the text, such as racism or sexism [197].

Moreover, Mei and Frank [198] present a semi-automated web-crawler for collecting extremist content using sentiment analysis. The system uses a decision tree that classifies the web pages into a set of classes by combining methods of web-crawling, parts-of-speech tagging, and sentiment analysis. The content is classified into one of three categories: content with extremist sentiment (pro-extremist class); news sources (neutral class); government or anti-extremist organisations (anti-extremist class); and content unrelated to extremism. Pitsilis et. al. [199] make use of a different approach to present a hate-detection scheme made up of recurrent neural network classifiers to distinguish racism and sexism from normal text taken from Twitter posts. Here, the authors harness a supervised deep-learning architecture for text classification in terms of hateful content, which incorporates features derived from the users' behavioural data, with particular regards to how the users' tendency to utter hate-speech, as expressed by their previous history, could leverage the performance of the model.

Other frameworks have been developed to supplement the findings of such hate-classification tools to mitigate the effects of online hate, including providing counter-narratives to directly address hateful content. For instance, Cuthberston et al. [200] present ParityBOT, a Twitter bot developed to counter abusive tweets aimed at women in politics by sending curated counter-speeches that support female political leaders. ParityBOT collects and classifies tweets directed at a known list of women candidates using Twitter's real-time streaming API and features extracted from previous hate-detection classifiers, including HATE Sonar. If a tweet is over a specified threshold of hatefulness, a positive tweet expressing encouragement or facts about women in politics would automatically be posted to inspire and uplift female politicians.

Similarly, Hatemeter [201] is a platform that makes use of a combination of NLP, machine learning, big data analytics, and visualisation techniques to identify

in real-time "red-flags" of anti-Muslim hate speech in order to understand and assess the sets of features associated with Islamophobia online. Insights gained from this are then used to form an effective strategy against anti-Muslim hatred using computer-aided persuasion approaches, and produce automated counter-narratives to address any identified hate speech.

In addition to the detection of hateful content, researchers have also developed frameworks with further specialised functionalities to explore features of hateful activity. This includes analysing the online networks formed by hate groups. For example, Chau and Xu [4] present an approach that harnesses semi-automatic techniques, to identify and analyse 28 anti-Black hate groups on the Xanga blog hosting site. This approach consists of four main modules: blog spider (downloads blog pages from the Web), information extraction (a pre-processing module to collect all textual content), network analysis (takes data about these blogs and their relationships to further analyse network structures), and visualisation (presents the analysis results to users in a graphical display). This is then used to provide insight into the structural properties of networks of blogs used by hate groups as well as identifying leaders of influence within them.

More recently, Uyheng and Carley [202] propose a dynamic network framework to characterise hate communities on Twitter, which they applied to discussions relating to the COVID-19 pandemic in the United States and the Philippines. The authors make use of various network-analysis techniques and features to assess how network clusters with higher levels of hate speech compare to non-hateful clusters in terms of structure and organisation.

Similarly, some frameworks explore these networks of hate in further detail to understand the dynamics of user interaction that facilitate the spread of hateful content within online social networks. Mathew et al. [21] propose a framework that looks into the content diffusion dynamics of posts made by hateful and non-hateful users on Gab. Their findings from using this framework suggest that content generated by hateful users tend to spread faster, further and reach a much wider audience as compared to content generated by normal users. Masud et al. provide

additional functionalities in their proposed framework, RETINA [203]. Here, the authors formalise the dynamics of hate generation and retweet spread on Twitter by analysing the activity history of Twitter users and signals propagated by the structural properties of networks on Twitter, which are generated by follower connections and events happening inside and outside of Twitter. The output from this is then used in RETINA to predict potential retweeters of a given tweet, thus modelling and predicting the diffusion dynamics of hateful content on Twitter. Such frameworks provide unique and valuable understanding into the spread of hateful content that is seldom explored within the context of online hate.

Despite all the extensive frameworks developed and applied to the study of online hate described above, the literature review conducted in this chapter showed that very few frameworks have been designed to explore how hateful behaviours and content compare and relate across different online platforms. One of the first frameworks to explore the wider network of online platforms was carried out by Zanettou et al. [29], as discussed previously. Here, the authors make use of network analysis and content diffusion analysis to create a statistical model to measure the influence that different online communities have on each other in news propagation. Similarly, few hate-detection frameworks make use of multiple platforms during both its development and evaluation. One of the few frameworks to do this was presented by Chandrasekharan et al. [183]. The "Bag of Communities" approach used in this study made use of training data from multiple platforms and online communities.

Through this comparison of existing online hate analysis frameworks, a number of gaps in the functionalities offered can be identified. Firstly, nearly all the frameworks examined focus on only a single data source or platform, making the framework dependent on a particular platform, such as Twitter. The analysis methods used are thus not generalised to other data sources or platforms. Since it has already been established that hateful activity often makes use of multiple platforms, a single-platform focused framework would fail at providing insight into the wider picture of online hate. Ensuring data from multiple platforms can be analysed with the framework would therefore be essential to gain a better understanding of hateful

activity online. Additionally, most of the frameworks made use of a single method of analysis, and thus focused on a single functionality or limited functionalities. This approach restricts the findings of the investigation as it provides a narrower view of a particular case study. Analysis frameworks should therefore harness a combination of analysis methods in order to offer a wider range of functionalities, and thus provide deeper insight into hateful activity online.

2.6 Summary

The literature survey presented in this chapter has provided an exhaustive exploration of the field of online hate, including approaches to gaining deeper insight into this phenomenon as well as providing potential solutions for it. Through this, gaps within the current research landscape have been identified, which have in turn informed the motivations for this project. Though it is apparent in this review that there are various unaddressed issues in tackling online hate or reducing its impact, it is particularly evident that an enhanced approach or methodology is required to gain a more complete understanding. More specifically, previous research in online hate has generally focussed around only one particular platform, even though there is sufficient evidence showing that hate groups often strategize the usage of different online platforms in order to circumvent current monitoring efforts, as has been discussed in the previous sections. Thus, this has resulted in a restricted, and at times unrealistic, understanding of this field.

Cross-platform analysis would therefore be an effective approach to address this gap by advancing and validating existing findings on online hate. Current research applying this approach for analysis is very limited, and generally provides preliminary insights. In order to extend this knowledge, any cross-platform analysis conducted must focus on a number of key aspects. Firstly, developing a social-media analysis framework requires clearly defined research objectives, selection of relevant data sources, appropriate data collection and analysis methods, consistent validation criteria, and adherence to ethical considerations. Following these requirements

would help to ensure the validity and effectiveness of the analysis framework and the research findings.

More specifically, developing an analysis framework with a cross-platform approach would entail the usage of data collected from a variety of different platforms. This cross-platform approach should also apply theoretical understanding from social-media studies to explore how hate groups with different ideologies, but at times overlapping interests, collectively advance narratives over time. Finally, it would be essential to model the wider structural dynamics of organised hate and how networks or clusters of hate are formed across multiple platforms, which would help investigate how content diffuses strategically through these interconnected networks.

Using this cross-platform approach will aid CSS researchers in gaining a more accurate image of the online global hate ecology, which could then inform how platform providers or law-enforcement agencies can potentially diminish its impacts, online and, in turn, offline.

3 Methodology

3.1 Introduction

The research conducted in this thesis is of a inter-disciplinary nature and utilises methods and concepts within computational social science (CSS). A mixed-methods approach was used to address the different research questions and the subsequent research objectives outlined in Chapter 1. In this chapter, a high-level description of the core research methods used throughout this thesis is provided. However, the detailed methodological steps followed are included in each subsequent chapter.

3.2 Data Collection

A variety of data sources are used in each of the studies conducted in this thesis. Primarily, this research focuses on applying the various analysis functionalities of the proposed framework to different case studies across multiple online platforms. In particular, the studies detailed in subsequent chapters analyse content from the platforms Twitter, Reddit, 4chan, and Stormfront, which were chosen specifically as they each offer distinct types of social-media platforms. Twitter is the largest and most mainstream platform of the four, and offers the largest and most diverse online audience. It is also the platform that moderates content the most, and therefore explicitly hateful content is often removed fairly quickly. Twitter data was

collected using the official Twitter API[1]. Reddit is another mainstream platform, though with a smaller audience size than Twitter. Content moderation is also carried out by Reddit, although not as much as Twitter. Here, hateful communities are often cultivated on particular subreddits, but such subreddits have also been removed from the platform if they are increasingly linked to hateful events, such as the subreddits r/fatpeoplehate and r/CoonTown [183]. The 4CAT Capture and Analysis Toolkit [204] was utilised to collect Reddit data from various subreddits.

Both 4chan and Stormfront, on the other hand, represent fringe communities with more specific audiences. 4chan is an anonymous imageboard platform with no content moderation, where online hate on the platform has been linked to several offline crimes and extremist attacks [11]. Again, 4CAT was used to collect data from the 4chan politically incorrect (/pol/) board. Stormfront is distinct from these platforms in that it prides itself in being "the first White Nationalist forum on the Web", where the platform actively tries to amplify white-supremacist voices and opinions. It has also been linked to hateful discourse that has resulted in several violent acts of extremism, including the mass killing of 77 people in Norway in 2011 [206]. Each of these four online platforms therefore provides a distinct set of functionalities and audiences. Stormfront data was retrieved from the "ExtremeBB" dataset provided by the Cambridge Cybercrime Centre[2] through a license agreement, and consists of a comprehensive collection of data from various extreme forums online, as detailed in [207].

Several key events have controlled online discourse around the world in the past few years including the the 2020 US presidential election and the COVID-19 pandemic. The 2020 US presidential election proved to be a key opportunity for various hateful narratives to be disseminated across several different online platforms. This is in line with previous research that has shown that election campaigns are a particularly conductive ground for breeding hate speech, to the point it has become a common emblem of political discourse [208]. Similarly, the COVID-19 pandemic not only brought a major international health crisis, but

[1]https://developer.twitter.com/en/docs/twitter-api
[2]https://www.cambridgecybercrime.uk/datasets.html

also highlighted inequalities and political polarisation, which oftentimes exposed and exacerbated conflicts between social groups [202]. The cross-platform analysis framework proposed in this thesis is therefore applied to these two case studies, which provides much needed insight into the discourse of hate ideologies and the impacts that various social issues and offline events have on them.

3.3 Computational Analyses

3.3.1 Natural Language Processing (NLP)

NLP is a field of Artificial Intelligence that gives machines the ability to read, understand and derive meaning from human languages. NLP consists of many different techniques for interpreting and manipulating human language, ranging from statistical machine learning methods to rules-based and algorithmic approaches [209]. A wide range of approaches is necessary as text and voice-based data varies widely, along with its practical applications. Basic NLP methods include tokenization and parsing, lemmatisation or stemming, part-of-speech tagging and identification of semantic relationships. Generally speaking, NLP involves the breaking down of language to shorter, elemental pieces, understanding the relationships between these pieces, and exploring how the pieces work together to create meaning. The main techniques used within this research include Bag of Words (BOW), TF-IDF, N-grams and topic classification. Further details of these techniques have been provided in Chapter 2.

Some of the applications of these techniques include content categorisation, topic discovery and modelling, contextual extraction and document summarisation. This research applies such NLP techniques to analyse hateful content retrieved from various platforms. In particular, these methods are used in Chapter 5 and Chapter 6 to understand the extent to which these platforms can play individual roles and serve different purposes within the wider ecosystem of online hate.

3.3.2 Machine Learning

Machine Learning can be defined as the process of automated detection of meaningful patterns in the data [16]. Machine-learning models take samples of labelled text to produce a classifier that is able to detect hateful content based on labels annotated by content reviewers. Machine-learning models are categorised as follows: supervised (labelled data) and unsupervised models (unlabelled data). A wide range of models using different machine-learning algorithms have been proposed with varying levels of accuracy in previous literature. Some of these algorithms include Support Vector Machines (SVM), Random Forests, Naive Bayes, logistic regression, and deep learning. This research will make use of machine-learning approaches to perform sentiment analysis and extract psychological context from data across multiple online platforms. Again, these methods are used in this research to understand how content on each platform is adapted to take advantage of platform affordances.

3.3.3 Social Network Analysis

Social Network Analysis (SNA), in general, studies the behaviour of the individual at the micro-level, the pattern of relationships (network structure) at the macro-level, and the interactions between the two through representing networks as graphs [210]. A typical social network representation has nodes for people, and edges connecting two nodes to represent the relationships between them. As the network representation of a particular community grows, it becomes necessary to apply graph analytic techniques to compute the characteristics of nodes and the graph as a whole [211]. Such analytic techniques are broadly focused on differences in centrality, investigating strongly connected clusters, and identifying positions that are structurally equivalent in networks, or of unique positions. Other measures enable the comparison of network structures as a whole, for instance, investigating content diffusion dynamics within the whole network. Such techniques are used in Chapter 6 to map the interaction of various hate groups or hate ideologies across multiple platforms, which is then used to model how hateful activity compares across different platforms.

3.3.4 Content Diffusion Analysis

Content diffusion analysis studies the process by which information is spread from one place to another through social interactions. This typically involves three main elements: *sender nodes* responsible for initiating the diffusion process; *receiver nodes*, which receive this information from the senders, where there are usually a greater number of receiver nodes in comparison to sender nodes; and the *medium*, which represents the channel through which the content diffuses from the senders to the receivers, such as a post on a social-media platform [212].

Content diffusion analysis typically builds upon findings from SNA, where the structural positions of nodes within a given network along with specific characteristics of these nodes and the content often determine the speed at which content diffuses through the network. Networks with different patterns of connected nodes have different properties regarding how content is propagated, which have significant implications for intervention measures [213]. These techniques for analysis are used in Chapter 6 to determine the extent to which content diffuses across platforms, further confirming that these platforms are part of larger networks of hate.

3.4 Validation of Framework

Through exploring the literature on online hate in the previous chapter, and comparing the frameworks developed for its analysis, it is apparent that there is not always a specific statistic or scale that can be used to measure online hate, or how well it can be analysed or mitigated. However, in order to reflect on the efficiency of the cross-platform analysis framework proposed in this thesis, approaches drawn from social sciences that are supported with the findings gained from computational analyses can be used to gain insight into the validity of the framework. As argued by Seale and Silverman, the quality of such research "cannot be determined by following prescribed formulas. Rather its quality lies in the power of its language to display a picture of the world" [214]. More specifically, this thesis makes use of case studies to assess the effects of using a cross-platform approach in online hate analysis.

Firstly, each of the functional components of the framework and the analysis methods included in these have been applied to various case studies to test their effectiveness at providing significant findings for multiple online platforms. This in turn allows for the framework to be assessed with regards to whether it provides any meaningful insights in comparison to an analysis approach developed for a single platform. This combined use of qualitative and quantitative methods can be used to support generalisations by particular events or case studies [215]. The findings from each of these analysis components of the framework are detailed in Chapter 5 and Chapter 6. The validation process also evaluates the performance and design of the framework against a list of validation criteria, which have been informed by the framework requirements listed in Chapter 4. The validity of the framework, including this list of validation criteria, are reflected upon in Chapter 7.

3.5 Ethical Considerations

Formal ethical approval was obtained from the Central University Research Ethics Committee at the University of Oxford. In addition to this, careful adherence to research-ethics guidelines were followed before, during and after this research, during the reporting of results, and during the storage of the collected data. This thesis primarily focuses on the overall online behaviours of individuals and groups driven by hateful ideologies. All of the data used in each study is publicly available, and can be viewed without having to create an account with any of the online platforms included in the research. Additionally, account handles, organisation names or quotations of posts from these platforms are not included in any of the studies that could be used to identify accounts, and therefore potentially people. Instead, only aggregate findings from the analysis of the posts are provided.

Any data supplied by other institutions, such as the Stormfront data provided by the Cambridge Cyber Crime Centre [207], has been used with license agreements in place to prevent any misuse, where the main purpose of collecting and using this data is to find, understand, investigate and counter political extremism. No data from these datasets are published in this thesis or any studies and, again, only

findings from the analysis of the data are included here. Furthermore, since this research involves the analysis of hateful and derogatory content, efforts have been made to censor any specific discriminatory terms throughout this thesis.

Finally, given the sensitive nature of the topic of this research, university guidance on research involving security-sensitive research material is followed, and measures have been put in place to consider the psychological safety of the main researcher.

3.6 Summary

In this chapter, a high-level overview of the methodology used to create the cross-platform analysis framework and apply it to subsequent studies has been presented. The usage of a range of quantitative and qualitative methods were justified, and an outline of the computational methods used to perform content analysis has also been provided. Finally, this chapter includes a brief discussion of the various ethical considerations associated with this research.

4

A Cross-Platform Analysis Framework

4.1 Introduction

Online hate is a complex phenomenon, with its definition varying across different theoretical paradigms, disciplines of study, and forms of victimisation [190]. Due to this complexity, online hate research is a fragmented field with a growing amount of research across various disciplines, as the adverse effects of online hate are more widely recognised in society. In addition to this, as society has advanced, the types of data available for analysis have expanded immensely as a consequence of new technologies, such as social media and artificial intelligence. Whilst this is highly advantageous for CSS researchers in terms of the vast amounts of data now available for study, this has highlighted the need for the development of computational methods to be able to handle and analyse large-scale, complex datasets efficiently, thus, research in this field has flourished — several of these research studies have been discussed in Chapter 2.

However, there is still a lack of accessibility for such researchers to large-scale analysis methods, oftentimes resulting in a barrier to exploring some of their research hypotheses. In order to aid researchers to better collect, analyse and understand hateful behaviours on online platforms, further study into the development of frameworks and tools for computational analysis is therefore required to lower

these skill barriers. Besides this, recent years have shown an increase in the number of online platforms introduced each year, as well as in the prevalence of users having multiple social-media accounts at any given time — with an average number of 7.2 platforms used each month per internet user [216]. In order for online-hate researchers to be able to gain more realistic understandings and insights into user behaviour, analysis frameworks must, consequently, also account for the usage of multiple data sources.

In this chapter, a novel framework for online-hate research that harnesses a cross-platform approach for analysis is proposed. This framework is built on a conceptual model of how various methods for analysis and prevention in online-hate research can be used to reduce the consequences associated with online hate. The requirements of the analysis framework that should be adhered to during its design and implementation are then defined and outlined. These requirements have largely been informed by the key findings and research gaps that were identified in the literature review carried out in Chapter 2. Following this, the structure of the framework, including the functional components comprising it, is presented, focusing on the specific methods used for the analysis. The various functionalities of the framework will then be applied to multiple case studies, the results from which are detailed in subsequent chapters of this thesis, and will be used in validating the framework.

The main contributions of this chapter are outlined as follows:

- Define the requirements of the analysis framework that have been informed by the findings from the literature review in Chapter 2.
- Provide a conceptual basis of how analysis and prevention methods can be used to mitigate the harms of online hate, and how this is considered in the development of the cross-platform analysis framework.
- Outline the design of the structure of the cross-platform analysis framework that details its main functionalities to aid online-hate researchers in analysing hateful content across multiple platforms.

4.2 Conceptual Model of Online Hate

In order to understand how the key functionalities for online-hate analysis frameworks discussed previosuly would help to counter or mitigate the risks of online hate, a conceptual model for online hate is presented in this section, as shown in Figure 4.1. This model includes the major components pertinent to online hate research: causes, consequences, and methods for analysis and countering. The purpose of presenting this model here is to provide an overview of research within each component. This model is then used to ensure that the potential impacts of each of the different components on online hate are considered during the designing and development of the cross-platform analysis framework proposed in this thesis.

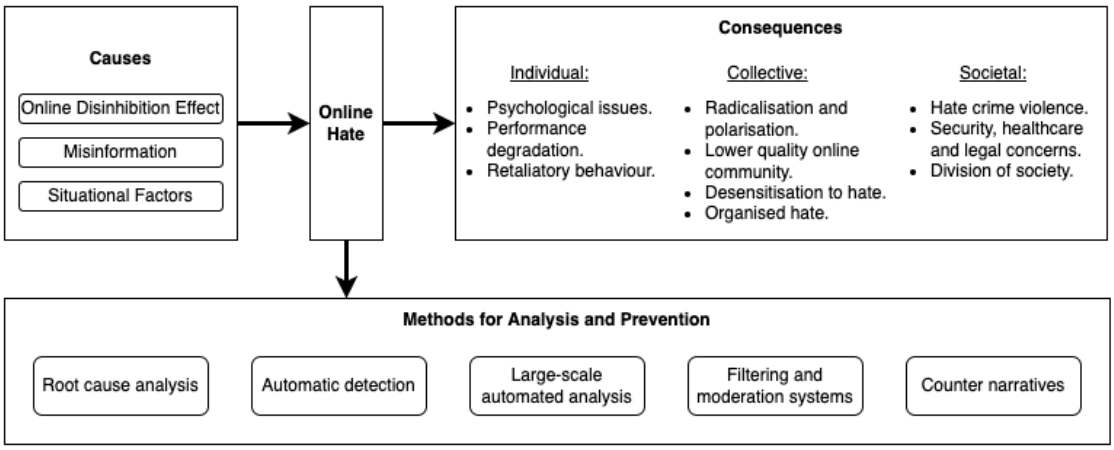

Figure 4.1: A conceptual model of online hate.

Along with the findings gained from the comparison of hate-analysis frameworks detailed in the literature review in Chapter 2, this conceptual model has also been used to help identify future directions of research and gaps in functionalities of current hate analysis-frameworks. The following sections discuss how each of the components within the conceptual model relate to each other, and should be addressed during the development or usage of the cross-platform analysis framework.

4.2.1 Causes of Online Hate

Online hate is a phenomenon that has arisen due to a combination of both technological advances and various social issues. Some of the root causes of online

hate are discussed here, though this is not an exhaustive list of all the possible causes. Instead, insight is drawn from academic literature to include the most prominent and most studied root causes. This includes the online disinhibition effect, misinformation as well as situational factors.

Online Disinhibition Effect. While online, some people self-disclose or act out more frequently or intensely than they would in person, especially anonymously [217]. This lack of restraint felt when communicating online in comparison to communicating in-person is known as the online disinhibition effect. In addition to anonymity, other factors that contribute to creating this effect include invisibility, asynchronicity, solipsistic introjection, dissociative imagination, and minimisation of authority [218]. This disinhibition effect could manifest in both positive and negative directions, which can thus be classified as either benign disinhibition or toxic disinhibition. Here, toxic disinhibition could lead to hateful, deviant or extremist behaviour online [219]. Several studies have found there to be a positive correlation between toxic online disinhibition and perpetrating both online hate as well as cyber-bullying [220].

Misinformation. The use of social media for peddling fake news has been highlighted on several occasions, including during elections [221], socio-political movements [222], and more recently during international health crises, such as the COVID-19 pandemic. Del Vicario et al. [154] carried out a large-scale quantitative analysis of the diffusion dynamics of fake news and conspiracy theories on Facebook. In their study, the authors observe that users tend to post and share content that adheres with their narrative and ignore other content, leading to the formation of echo chambers. When the primary driver of these echo chambers is fake news, it can cause polarisation, mistrust, conspiracies and paranoia [223]. This polarisation and paranoia often manifests itself as online hate when it is targeted towards an individual or group. This has been demonstrated over the course of the COVID-19 pandemic, most prominently with anti-Asian rhetoric [224]. In their study of analysing comments on Italian YouTube videos related to COVID-19, Cinelli et al.

[225] also show that users skewed towards more questionable content were more prone to using inappropriate, violent or hateful language.

Situational Factors. Various situational factors have also been shown to cause and exacerbate online hate. Cheng et al. [226] observe how factors such as personal mood, discussion context and the "contagiousness" of content manifest hateful activity. They find that personal dissatisfaction, bad moods and anger increases aggression towards others, which can lead to malicious behaviour online [226]. According to their study of trolling behaviour on online platforms, the immediate context of the discussion can hold the direction of the conversation, thus a single malicious user or post can lead to multiple users engaging in the proliferation of hateful content. In addition to this, several studies have found that different social traits can also determine the likelihood of a user participating in or responding to hateful content. These traits include: compliance, where the more compliant an individual is, the more receptive they may be to hateful content, especially from self-proclaimed figures of authority [227]; naivety, since they would be more likely to conform to others around them [228]; as well as emotional vulnerability and social deprivation [229, 230].

4.2.2 Consequences of Online Hate

As has been discussed in previous chapters, online hate has been linked to several abhorrent consequences, including offline violence and extremist attacks, as well as psychological problems and emotional trauma. The consequences included within this conceptual model are loosely categorised based on the frameworks developed by Salminen et al. [43] and Chaudhary et al. [231], where the harms are divided into three types: individual, collective and societal harms.

Individual Harms. Individual harms refer to the consequences that individuals face from online hate. Previous studies have explored how online hate can lead to several negative psychological effects on individuals, including depression, anxiety, socio-psychological problems, and has even been linked to an increase in suicides

[10]. Other works have discussed how it can lead to performance degradation at work or school, and promoting self-harm as well as retaliatory behaviour [220].

Collective Harms. Collective harms describe the consequences faced by a group of individuals. On a collective level, it is often argued that online hate can lead to radicalisation [232], group polarisation where the previously held prejudices are enforced [154], degraded quality ("health") of an online community [119, 233], and decreased feelings of safety and well-being of online users [234]. It can also cause desensitisation to hateful statements in online users and lead to organised hate-speech against marginalised groups [235, 236].

Societal Harms. Societal harms deal with the harms caused to the wider society. The manifestation of increasingly polarised online groups can lead to divides in society due to the collective trauma experienced by the targeted group within online environments. In extreme cases, it can also result in offline violence and extremist attacks [237, 238], which add to security, healthcare and legal concerns.

4.2.3 Methods for Analysis and Prevention

As discussed in previous chapters, online hate research has introduced a plethora of methods for its analysis and prevention. Oftentimes, multiple approaches are combined within analysis tools and frameworks. Much of these are centered around the automatic detection of hate speech, which have generally been used within filtering and moderation systems. Machine learning and artificial intelligence are increasingly being deployed to fill important moderation functions on various online platforms. Incidents like the 2019 Christchurch terror attack clearly show that automated moderation systems have become necessary to manage growing public expectations for increased platform responsibility, safety and security [239].

Following major public controversy regarding Facebook's role in increasing hate towards Rohingya refugees in Myanmar [240], Facebook improved its Myanmar-language hate-speech classifiers, leading to a 39% increase in automated removals [241]. Similarly, YouTube reports that "98% of the videos removed for violent extremism are flagged by machine-learning algorithms", with Twitter also reporting

that, of the thousands of accounts suspended for promoting terrorist propaganda, 93% were flagged using automated detection mechanisms [239, 241]. This demonstrates that such approaches have become integral to the operation and governance of most platform providers.

Advances in machine-learning and artificial-intelligence methods have not only been used in automated detection and moderation systems, but have also allowed for researchers to carry out large-scale analysis of hateful content. This has, in turn, facilitated the understanding of the types of content promoted by hateful users, in addition to gaining deeper insight into their online activity. Such methods have been used in previous studies to distinguish between different types of abuse and the degree to which it is explicit [242], as well as to study the relationships of hate instigators and targets [243].

Large-scale analysis has also allowed further exploration of the group dynamics between hateful users and organisations, including radicalisation and persuasion techniques [232, 244], the cultural transmission of hate [245], as well as social exclusion [246]. Many of the insights gained from this are often used in combination with various theoretical frameworks from social science on group behaviour [57], persuasion [232], and motives for perpetuating hate [247] to understand the root causes of online hate.

Recent practices in online-hate analysis and prevention have introduced the usage of counter-narratives to complement other analysis and prevention methods. For instance, Cuthberston et al. [200] make use of counter-narratives in their tool, ParityBOT, which combines automated detection of hate speech targeted at female politicians on Twitter with an automatic counter-speech generator. Chung et al. make use of root-cause analysis as well as some large-scale analysis techniques to provide a multilingual dataset of hate speech/counter-narrative pairs [248]. Here, root-cause analysis was used to understand the types of counter-narratives that would be most effective for countering different types of hate speech.

4.3 Requirements for Designing a Cross-Platform Analysis Framework

In addition to establishing a conceptual model of online hate, it is necessary to define and adapt certain requirements before designing a cross-platform analysis framework for the study of online hate. Through reviewing existing literature from the field of social-media analysis, a list of requirements that any analysis framework should aim to adhere to was presented in Chapter 2. These requirements provide the key factors that must be considered when developing a framework for collecting and analysing data from social-media platforms.

Since online hate is prevalent across a variety of online platforms, adapting these requirements to include a cross-platform approach to online hate research can help researchers gain a more comprehensive understanding of the nature and prevalence of hate in online spaces. Ensuring that these requirements have been tailored to the study of online hate also ensures that the necessary ethical considerations are taken into account. Due to the significant impacts that online hate can have on individuals and communities, it is important to minimise harm in the research process. Adapting the framework requirements to include ethical considerations, such as obtaining informed consent and anonymising data where possible, makes certain that research is being conducted in a responsible and ethical manner.

The requirements for developing an analysis framework, which are initially outlined in Chapter 2, have been modified as follows to determine the final list of requirements that have then been used in the designing of the cross-platform analysis framework for the study of online hate proposed in this thesis:

1. *Clearly defined research objectives:* within the context of this thesis, the objectives of the cross-platform analysis framework are largely aligned with the research questions and objectives defined in Chapter 1. Namely, the framework should provide novel insight into the cross-platform behaviours of hateful groups and users, so as to gain a better understanding of the prevalence and impacts of online hate.

2. *Selection of relevant data sources:* for this analysis framework, relevant data sources include social-media platforms and online forums that are known to host hate speech. In order for this framework to be designed with a cross-platform approach, it is required for more than one data source to be used with the framework. For instance, within this thesis, the data analysed with this framework is extracted from the platforms Twitter, Reddit, 4chan and Stormfront. These platforms were chosen as they each offer distinct types of social-media platforms, including both mainstream and underground — further details are provided in Chapter 3.

3. *Appropriate data collection methods:* the methods used to collect data from the relevant data sources within this analysis framework include the usage of platform APIs, social-media analysis toolkits, and datasets published by other academic researchers. Data is collected using filtering techniques from hate-specific users or environments; again, more details are provided in Chapter 3.

4. *Valid and reliable data analysis methods:* this cross-platform analysis framework utilises various concepts and methods from CSS. This includes NLP methods (such as topic discovery and analysis), machine learning techniques (such as sentiment analysis), as well as social network analysis methods (such as URL co-occurrence and domain network analysis). After reviewing previous literature within the field of online hate, these particular methods were selected for the novel insights they would provide through a cross-platform perspective.

5. *Consistent validation criteria:* the validation criteria used within this analysis framework should be consistent with the research objectives of this thesis. This also involves the usage of multiple case studies to assess the effects of using a cross-platform approach in online-hate analysis — further details on this as well as the validation criteria are provided in Chapter 7.

6. *Ethical Considerations:* it is essential to ensure that the relevant ethical guidelines and approval boards have been consulted prior to carrying out any

part of the cross-platform analysis. This is to ensure that data is anonymised where necessary, and data protection and privacy laws are adhered to at all times. Chapter 3 discusses the ethical considerations and approvals that were consulted within this thesis.

By adapting these requirements of social-media analysis frameworks to a cross-platform approach for online-hate analysis, researchers can gain a more comprehensive understanding of the nature and prevalence of online hate in the wider hate ecosystem. Adhering to these requirements in the designing of the analysis framework provides better structure and guidance during any analysis carried out, to ensure that the collected data is comprehensively studied and correctly interpreted.

4.4 Structure of the Framework

One of the challenges that researchers — particularly those from a non-computational background — face when investigating online hate is making sense of large amounts of data. As such, the structure of the cross-platform analysis framework is designed to assist researchers in handling complex data and supporting multiple analytical techniques. The results from any analysis carried out using this framework are then used to draw insights into cross-platform activity in online hate. An overview of the architecture is shown in Figure 4.2.

4.4.1 Framework Components

The cross-platform framework consists of three main layers: a data-sources layer, a methods layer, and an analysis layer. The data-sources layer requires for data from more than one online source that are of interest to the researcher to be used with the framework, where this data is then fed into the methods layer for analysis. Using multiple data sources here is essential to the main function of the analysis framework to gain any understanding of cross-platform activity.

The next layer is the methods layer, which provides a collection of analytical functionalities to support the analysis of the uploaded data. Methods such as

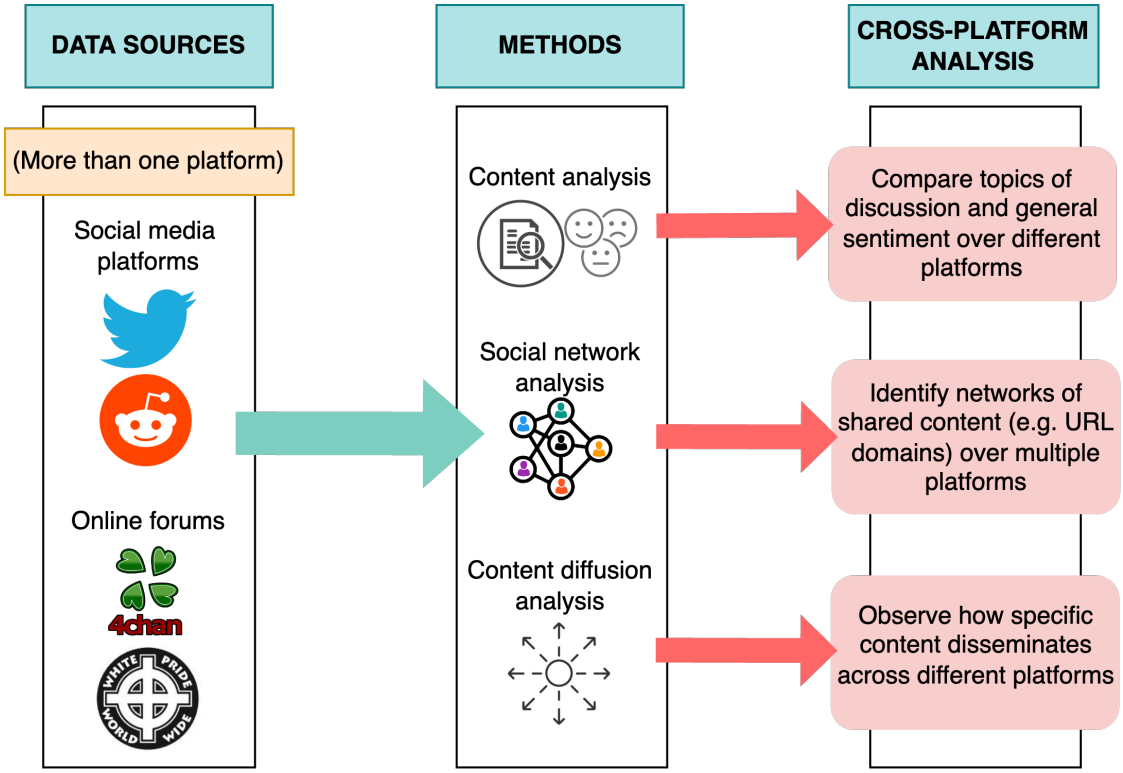

Figure 4.2: Structure of the cross-platform analysis framework.

content and sentiment analysis, psychological analysis, social network analysis, and content diffusion analysis are used to perform these operations. These methods are utilised to support different functionalities that a researcher may need to perform.

Following on from this is the cross-platform analysis layer, where the insights gained from the analysis previously are used to identify and understand cross-platform behaviours and activities across the platforms being analysed. This includes using the various functionalities to compare topics of discussion and sentiment over different platforms, identifying networks of shared content over the different platforms, as well as observing how specific content disseminates across them.

4.4.2 Analysis Tasks and Methods

Based on the comparison of previous analysis frameworks for online hate presented in academic literature, as detailed in Chapter 2, as well as the conceptual model of online hate provided in the previous section, a variety of methods used in the analysis of online hate were highlighted. Some of these methods have been used

more frequently within the research landscape, particularly automated detection. However, other methods, such as social network analysis and content diffusion analysis have rarely been applied to cross-platform studies of online hate. As such, the main analysis tasks of the proposed cross-platform analysis framework and the methods used for these tasks are outlined below.

Content Analysis

The content analysis task allows online-hate researchers to analyse the content posted by hateful users. This method utilises techniques from natural language processing, data mining, and machine learning. For example, entity extraction is used to allow investigators to extract specific patterns such as URLs. Additionally, this functionality allows the researcher to analyse the data in order to identify the main topics of discussion, explore the linguistic composition of online posts, and compare the sentiment and psychological motivation of content across multiple platforms.

Social Network Analysis (SNA)

The SNA task allows researchers to identify and capture interactions between different users within a hateful network. For example, it can be used to create a network of possible hateful "nodes" and the interactions or "edges" between them. Using SNA techniques, researchers are able to detect any cross-platform activity and communication, to gain a deeper understanding into how multiple platforms are utilised in wider networks of online hate, and provide visual representations of hateful networks. Additionally, this functionality provides insight into the influence of particular platforms as well as the roles they may play in perpetuating online hate. Through the findings from this, researchers may be able to identify potential ways in which these networks can be disrupted.

Content Diffusion Analysis

Content diffusion analysis forms the final functionality of the cross-platform analysis framework. This particular functionality is designed to complement the findings from the SNA task described above. Here, the content diffusion dynamics of online hate

are analysed to provide some insight into how content disseminates across platforms and within the hateful networks previously identified. Given the ever-increasing amount of user behaviour and interaction provided by social-media platforms, the paths traversed by content through online networks have become more complex, facilitating the alteration and adaption of content as it spreads. This analysis of the content diffusion dynamics would thus also provide novel insight into how content is presented and adapted to particular platforms within wider networks of hate.

4.5 Summary

In this chapter, a cross-platform analysis framework for online hate is presented based on the findings gained through investigating academic literature within this field. A conceptual model of online hate has been presented to provide insight into how various methods for analysis and prevention are utilised to address the causes and consequences of online hate.

In addition to this, this chapter defines the requirements that were used to guide the designing of the analysis framework. These requirements were largely informed by the findings gained from carrying out the literature review in Chapter 2, where the key functionalities and considerations of any social-media analysis framework were outlined. These were then adapted to align with the research objectives of this thesis, and address the major research limitations found in the literature review through the development of a cross-platform analysis framework for online hate.

The insights from both the conceptual model and the requirements were then used to design the structure of the analysis framework. As well as providing novel perspectives into the usage of multiple platforms in online hate, this framework has been developed to ensure that combined methods, such as content and sentiment analysis, social network analysis, and content-diffusion analysis, were prioritised in the framework functionalities, so as to address current gaps within academic literature.

The following chapters apply this cross-platform analysis framework to various case studies relating to online hate, with a particular focus on hate perpetuated from white-supremacist ideologies.

5

Determining How Different Platforms are Used in Online Hate

5.1 Introduction

The concept of online hate is still considered a complex phenomenon with an ever-evolving definition, thus, research into online hate is fragmented across numerous disciplines. One key area of study within this field is the relation between offline events and online hate, where the majority of the research exploring this is based on case studies. For instance, Burnap et al. perform a quantitative case study of the social media reaction after the Woolwich terrorist attack in the United Kingdom in May 2013 [17]. Similarly, Müller and Schwarz further empirically analyse the relation between online hate and hate crimes against refugees in Germany by performing fixed-effects panel regression on data retrieved from 2015 to 2017 [249], where they find that social media often acts as a propagation mechanism for violent crimes by enabling the spread of extreme viewpoints.

Despite all the extensive approaches proposed to analyse online hate within the research landscape, limited studies have investigated how hateful behaviours and content compare across different online platforms [12, 26, 181]. Although research within this aspect of online hate is scarce, in the last few years, a few studies have realised the importance of the insights that can be gained from cross-platform

analysis. With this motivation, Phadke and Chandaluri conduct a preliminary study where they collect data from the Twitter and Facebook accounts of various hate groups, and explore how content is framed and shared across both platforms [26]. Through this, the authors highlight some differences in the way both the platforms were used by hate groups, where Facebook seemed to be used for group radicalisation and recruitment, and Twitter was mainly used to reach a diverse follower base.

More recently, Hitkul et al. [250] conducted a comparative study of Twitter and Parler content during the aftermath of the 2021 Capitol riots. Though this study was not focused on hateful content, it still provides some insight into how sentiment and narratives can differ across platforms. Similarly, Murdock et al. conduct a multi-platform study of fraud and protest-related posts on Twitter, Facebook and Reddit in the aftermath of the 2020 US election [251].

This chapter will aim to build on this particular line of research by harnessing the cross-platform analysis framework developed in Chapter 4 to gain a clearer understanding of the dynamics of the global hate ecosystem. In particular, this study will make use of data collected over the course of the 2020 US presidential election and the COVID-19 pandemic from four different social-media platforms — Twitter, Reddit, 4chan and Stormfront — to investigate how hateful content and narratives compare across multiple platforms. This research builds on previous work by exploring both mainstream platforms, such as Twitter and Reddit, as well as non-moderated fringe platforms, like 4chan and Stormfront; more details on the data collection process are given in Section 5.2.1 as well as Chapter 3.

More specifically, this chapter details the application of various computational methods, including topic modelling, linguistic analysis and sentiment analysis, to explore the type of content that is promoted on each platform. This is conducted with the aim to gain some understanding on how online platforms are used for the different functionalities they offer, and how specific platforms can play a different role within the greater hate ecosystem. Thus, the findings detailed in this chapter aim to further fill the gap currently within this research landscape by providing more extensive empirical and statistical insight into the cross-platform behaviours of

online hate on both mainstream and fringe communities, within the context of the 2020 US election and the COVID-19 pandemic. Some of these findings are discussed in [12]. This provides some understanding into the type of content promoted on each platform and the linguistic composition of their posts.

The contributions of this chapter are as follows:

- Data from four different online platforms (Twitter, Reddit, 4chan and Stormfront) is collected over the course of the 2020 US election and during the peak of the COVID-19 pandemic between 2020 and 2021. The participation trends during these collection periods are also analysed.
- This research further conducts topic modelling to show how different types of content and narratives are promoted on each platform.
- A deeper study into the linguistic composition of the posts from each platform is carried out, and distinctions in the type of sentiment and level of emotion used are identified.

The remainder of this chapter is structured as follows. Section 5.2 provides a detailed account of the approach and methodology, including the datasets and data-analysis tools that were used. The results and observations from the findings are then discussed in Section 5.3 and Section 5.4, where findings from the analysis are presented for each case study.

5.2 Methodology

This cross-platform analysis of online hate during the 2020 US election and COVID-19 pandemic on Twitter, Reddit, 4chan and Stormfront is largely focused on content from white-supremacist ideologies, and is carried out with particular regard to the following research aims:

- **RA1:** Investigate how the participation and posting trends compare across all four platforms over the course of the election and COVID-19.

- **RA2:** Identify the main topics of discussion for hateful users and how they compare on each of the platforms.

- **RA3:** Identify any similarities or differences in the linguistic composition or general sentiment of the posts from the four platforms.

The approach used in this content analysis thus comprises three stages: (1) observing the posting behaviours of the data collected from all four platforms, (2) conducting topic modelling on each corpus of posts from the four platforms, and (3) carrying out a more in-depth linguistic analysis of the collected posts to examine their structural properties. Further details on the methods that were used at each stage are provided below.

5.2.1 Data Collection

As mentioned previously in Chapter 3, the Twitter datasets were collected using the official Twitter API, thus in order to adhere to rate and collection limits, further filtering methods were used. Since Twitter is a more moderated platform in general, it may be argued that only a small percentage of content will be identified as hateful (or the period in which hateful content is present on Twitter will be comparatively less than other platforms). To ensure only hateful content from predominantly white-supremacist users were selected, tweets were collected from the accounts of white-supremacist groups and their supporters.

These hate groups were identified from a list of hate groups published by the Southern Poverty Law Center (SPLC)[1]. This list contains approximately 300 hate organisations from various ideologies, of which 84 are groups supporting white supremacy; it should be noted here that the SPLC identifies these groups as white nationalist, neo-Nazi and neo-confederate, but these groups were combined in this research since they have shared views on extreme-right ideology and reported hatred for other races [26]. From these 84 hate groups, the associated Twitter accounts of 48 groups were found. A two-step snowball-sampling approach was then used

[1]https://www.splcenter.org/hate-map

to identify other hateful groups and users from the follower and followee lists of these accounts. Through this, 478 hateful Twitter accounts were identified, from which tweets relating to the US election and the COVID-19 pandemic were collected over the course of the respective collection periods.

The 4CAT Capture and Analysis Toolkit was used to collect the relevant data from both Reddit and 4chan. The Reddit posts were collected using the pushshift API [13] from the r/donaldtrump subreddit, which was linked to spreading online hate over the course of the election and pandemic, and was consequently banned by Reddit in the aftermath of the Capitol riots [252]. Similarly, 4chan posts were collected from the Politically Incorrect (/pol/) board, which has also been identified as a key platform for spreading online hate, and has been linked to several violent acts of extremism including the 2019 Christchurch shooting [11].

For the Stormfront datasets, the "ExtremeBB" dataset provided by the Cambridge Cybercrime Centre was utilised, which is a comprehensive collection of data from various extreme online forums. This research only made use of the collection of Stormfront posts from this dataset, which were further filtered to include only content posted over the course of the respective collection periods. After collecting all the datasets, the participation trends were measured by observing the frequency of posts being shared online over the course of the election and COVID-19, in answer to RA1. The results from this are discussed in Section 5.3.1 and Section 5.4.1. Further detail on the collection periods for the datasets and the filtering methods used to collect only relevant data is provided below.

US Election Data

In order to gain a comprehensive view of the online discourse over the course of the US election from all four platforms, data was collected from 1st October 2020 to 31st January 2021. This time frame encompassed all the key events that strongly influenced much of the online discussions in the lead up to the election and in the immediate aftermath, including the October presidential debates, the actual election date in November, as well as the electoral certifications and the subsequent

Capitol riots that took place in January. To collect this data, a list of key terms related to the US election and white-supremacist movements was used to ensure that only content related to the election was collected. The terms used are: “vote”, “election”, “maga”, “make america great again”, “trump”, “biden”, “democrat”, “republican”,“proud boys” and “stop the steal”. All four of the datasets were filtered based on these terms during the collection process.

The sizes of the four datasets are as follows:

1. Twitter Election dataset: 1,498,154 posts.
2. Reddit Election dataset: 112,981 posts.
3. 4chan Election dataset: 1,086,053 posts.
4. Stormfront Election dataset: 96,254 posts.

COVID-19 Data

With the aim of capturing a comprehensive overview of the online conversations across all four platforms during the COVID-19 pandemic, data was collected from January 2020 to March 2021. A majority of all the significant events that had a profound impact on the online discourse related to the pandemic were included within this time frame. More specifically, COVID-19 was declared as a pandemic by the World Health Organisation in January 2020 [253], and was still very much ongoing as of March 2021, where the roll-out of the COVID-19 vaccines was well underway, though some lockdowns and restrictions were being lifted [254]. This time frame for data collection therefore captures the majority of the duration of the pandemic, including its early stages and its ongoing impact. Again, to ensure that only content related to COVID-19 was collected, a list of key terms related to the pandemic and white-supremacist movements was used to filter content during data collection. The terms used in this case are: “covid”, “corona”, “pandemic”, “virus”, “lockdown” and “mask”. All four of the datasets were filtered using these terms during the collection process.

The sizes of the four datasets are as follows:

1. Twitter COVID-19 dataset: 1,361,580 posts
2. Reddit COVID-19 dataset: 46,977 posts
3. 4chan COVID-19 dataset: 845,982 posts
4. Stormfront COVID-19 dataset: 12,281 posts

5.2.2 Identifying Topics of Discussion

In order to identify the main topics of discussion during the course of the US election and the COVID-19 pandemic, so as to address RA2, topic modelling is conducted on each of the datasets. Using the Latent Dirichlet Allocation (LDA) topic detection model proved to work better overall on all the datasets than other models, like Non-Negative Matrix Factorization (NMF), even though NMF usually works better with shorter texts [255]. LDA topic modelling has been used to identify topics within social media posts in many previous studies [256, 257], and works under the assumption that a document is comprised of a collection of latent topics [258]. The model uses probabilistic assignments of terms to a user specified number of topics. From this, each unique term in the corpus is assigned a probability distribution relative to the number of topics, indicating for each topic the probability that the term occurs within it, thus providing a distribution of topics over documents.

As LDA topic modelling requires a user-specified number of topics, the topic model was experimented with different numbers of topics across each of the datasets [259]. From this experimentation, the number of topics that produced the most distinct topics in all the datasets was found to be five, thus this is the final number of topics identified in each of the datasets in the topic analysis. This work further assesses the extent to which each topic is discussed in every dataset, where the dominant topic in each post is found, and then the proportion of posts containing reference to that topic within the overall dataset is extracted.

A series of pre-processing steps were carried out before the linguistic analysis to clean and prepare the posts in each dataset (the datasets were not pre-processed when observing the frequency of posting over the course of the election and COVID-19). These steps included: (1) Removing any duplicate posts from the datasets to

reduce the levels of noise. (2) Removing all punctuation marks. (3) Removing any URLs (though these were kept separate for further analysis). (4) Removing any short posts (those less than 5 tokens). (5) Removing any platform-specific noise, for instance 'RT' for the Twitter dataset. All of the posts were then tokenized and a term-frequency inverse-document frequency (TF-IDF) array was created to fit the LDA model, which has been suggested by previous work to yield more accurate topics [260]. This analysis is carried out using the Pandas[2] data-analysis library and the 'Natural Language Toolkit' (NLTK)[3] provided by the Python programming language, where the LDA topic modelling was conducted with a Gibbs sampler using the Python Gensim wrapper.

5.2.3 Analysing Linguistic Compositions

In order to address RA3 and further linguistically analyse each of the datasets, the programmatically coded dictionary from the Linguistic Inquiry and Word Count (LIWC 2015) [261] analysis tool is used to automate the process of extracting further information on linguistic structures and psychological meaning from textual content. LIWC is a widely used tool in lexical approaches for personality measurement, and statistically analyses textual content based on 81 different categories by calculating the mean percentage of words in the input text that match predefined words in a given category [261]. Many previous studies from various disciplines have utilised LIWC to gain a more in-depth understanding of the structural and functional constructs used within language [262, 263], as well as to get insight into the psychological meaning of textual content [130].

LIWC is a text analysis software that examines written or spoken language, and categorises words into various linguistic and psychological dimensions. LIWC uses a pre-defined dictionary that contains words and phrases classified into different categories, such as emotions, cognitive processes, social processes, and more. It then calculates the frequencies of words falling into each category within a given

[2]https://pandas.pydata.org/
[3]https://www.nltk.org/

text and provides statistical insights into the text's content. More specifically, this statistical analysis involves the following steps:

1. LIWC comes with a built-in dictionary containing words and phrases grouped into various linguistic and psychological categories. For instance, it has a category for positive emotions, negative emotions, pronouns, cognitive processes, and so on.
2. The input text is divided into individual words or tokens. Punctuation, numbers, and special characters are typically removed or ignored during this process.
3. Each token from the text is compared against the LIWC dictionary. If a match is found, the token is associated with the corresponding category or categories from the dictionary. Some words might belong to multiple categories.
4. After categorising the words in the text, LIWC calculates the frequency of words in each category. This involves counting how many words from the text fall into each category. This step provides a numerical representation of the linguistic and psychological features present in the text.
5. The raw frequency counts are often normalised to account for differences in text length. This normalisation allows for fair comparisons between texts of varying lengths.
6. Once the text has been processed and the frequency counts have been calculated and normalised, LIWC can generate various statistical insights. These insights can include the proportions of words in different categories, comparisons between texts, and correlations between linguistic categories and psychological constructs.
7. Researchers can then interpret the results to gain insights into the emotional, cognitive, and social content of the analysed text. For example, they might identify patterns in the use of positive and negative emotions, cognitive processes, or the use of specific pronouns.

It is worth noting that LIWC's analysis is based on patterns in word usage, and it doesn't capture nuances such as tone, context, sarcasm, or idiomatic expressions. While LIWC can provide valuable insights, human interpretation is often required to understand the full meaning and context of the analysed text. LIWC is used in this approach to both examine and compare the functional composition of the posts collected from each platform, as well as to extract psychological meaning and sentiment from the datasets.

To do this, each dataset of posts is analysed with all 81 LIWC categories. This analysis focuses particularly on the four summary linguistic variables ('analytical thinking', 'clout', 'authenticity', and 'emotional tone'), and 10 more detailed variables that reflect the psychological states, linguistic dimensions, personal concerns, and informal language within each dataset. More specifically, this involves the usage of pronouns ('i', 'we', 'you', 'they') as well as emotive language, which used the LIWC categories 'positive emotion', 'negative emotion', 'anger' and 'anxiety'. These are measured by dictionaries of words associated with each category. This is used in the cross-platform analysis framework to identify the types of narratives that are promoted on each platform, as well as to gain further insight into the target audience that each platform addresses.

5.3 Case Study 1: 2020 US Election

5.3.1 Participation Trends

With each dataset, the analysis first explores the frequency of content being posted over the course of the collection period during the US election, in answer to RA1. In Figure 5.1, it can immediately be seen that the amount of content posted on Twitter is much greater than the other platforms. The number of posts on 4chan relating to the election is also considerably high, whereas it is clear that content is posted much less frequently on Stormfront. Somewhat similar trends in the amount of participation can still be seen, though, across all platforms over the course of the election time frame. This is highlighted in Figure 5.1, where a graph with the normalised data from all four platforms is included, using standard score

normalisation [264]. Here, post frequency is measured weekly, so as to examine how offline events related to the election affect online behaviour.

Notably, a significant peak in the number of posts can be seen in the weeks during the election (in the week beginning November 2nd), and again in the first week of January following Joe Biden's presidential certification and the resulting Capitol riots [265]. It is also worth noting that the Twitter dataset surprisingly exhibits this second peak in the lead up to the riots and during them (at the end of December and in the first week of January), whereas this peak appears in the 4chan dataset in the aftermath of the riots (in the last few weeks of January, which could be related to the presidential inauguration on January 20th).

In contrast, this peak is not as prominent in the Reddit dataset. It can also be observed that, although some small peaks can be identified during the two mentioned events, Stormfront has the most steady posting behaviour out of all four platforms, which consist of several peaks and drops in the frequency of posting. The r/donaldtrump subreddit was banned as a result of the part it played during these events, which is why the posts abruptly stopped in the last few weeks of the data collection time frame. Therefore, in answer to RA1, the participation trends across all four platforms seem to be similar, in that they generally have two major peaks at around the same time frames following key events over the course of the election.

5.3.2 Keywords and Topic Analysis

The cross-platform analysis next determines which words and topics were mentioned the most on each platform. The findings from this are, in turn, used to address RA2 by identifying the main topics of discussion on each platform. The word clouds shown in Figure 5.2 demonstrate that the most used key terms on all platforms mainly include *election* and *Trump*, with 4chan and Reddit also mentioning *Biden*. This could, in part, be due to the terms used to filter posts during the data collection, however these were applied to all four datasets and the filtering was just to ensure that only content related to the election was collected.

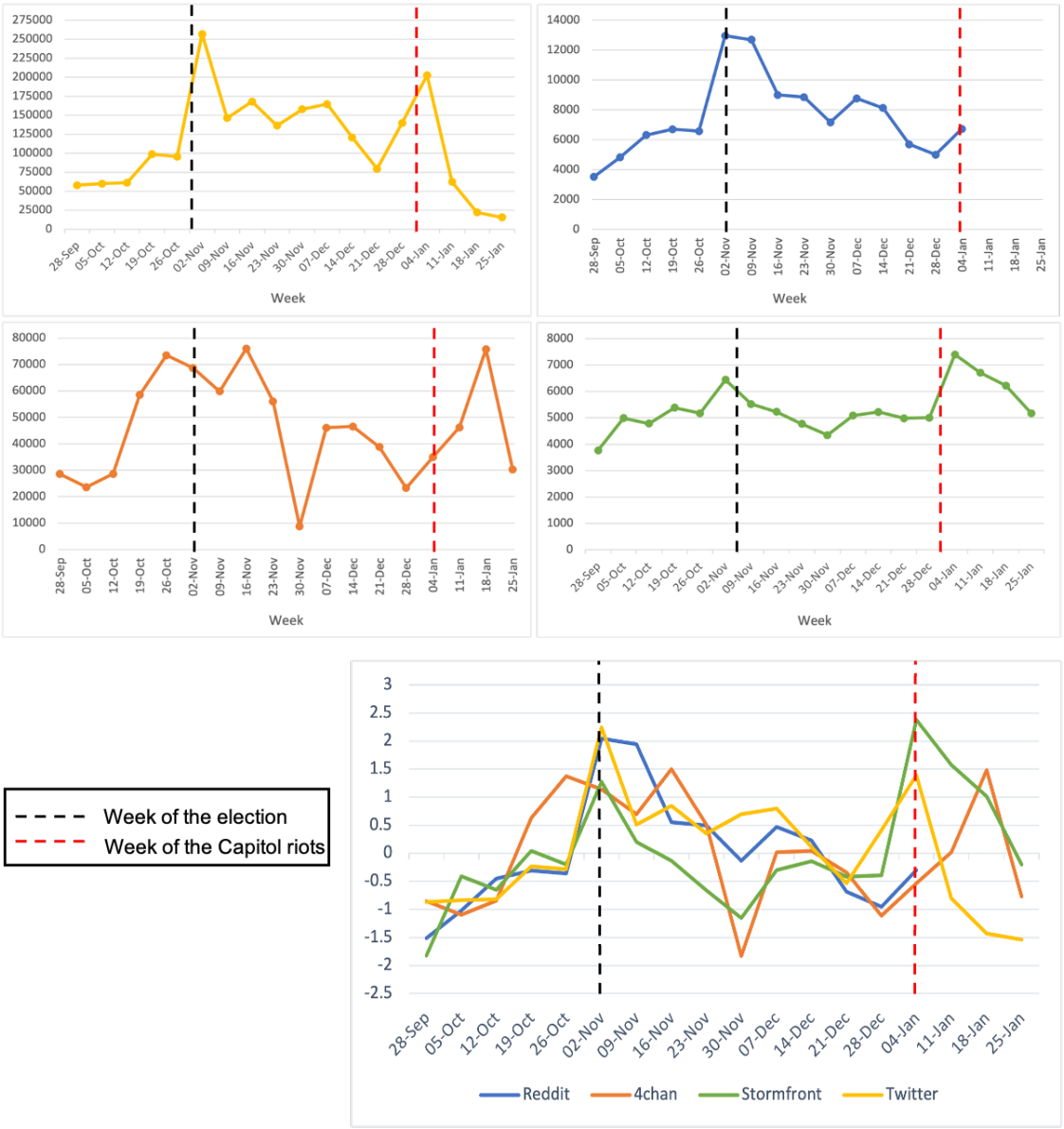

Figure 5.1: Graphs showing the frequency of posts across each dataset over the course of the 2020 US election: Twitter (top-left), Reddit (top-right), 4chan (bottom-left), and Stormfront (bottom-right). A graph with the normalised data from all four datasets is shown at the bottom.

In the 4chan dataset, it is clear to see that the posts often refer to groups of "others", such as *jew*, and use offensive and hateful names for them. Overall, a lot more explicit and derogatory language is used on this platform in comparison to the others. This could largely be attributed to the fact that this platform is known for its lack of moderation and the freedom it gives to users to post without fear of repercussion. This is consistent with previous research, which has established that 4chan has a culture of trolling and shock value, which has led to a normalisation

of hate speech and discriminatory language on the platform [266].

The word cloud from the Stormfront dataset shows that words such as *white*, *jew* and *black* were also used frequently on this platform. Previous research has shown that this platform has been known for promoting white supremacist and neo-Nazi ideologies, as well as anti-Semitic, anti-black, and anti-immigrant views [267]. Given the focus on race and ethnicity in these narratives, it is not surprising that the terms "white", "black", and "jew" would be mentioned most frequently on Stormfront. The site is dedicated to promoting white supremacist views, and so the term "white" is central to the platform's ideology. On the other hand, the word cloud from the Twitter dataset suggests content posted here seemingly mentioned *voter fraud* and *stop the steal*, which were common terms and slogans used by Trump supporters after the results of the election, where they claimed that the election was not conducted fairly [251]. Other commonly used slogans such as *proudboy* and *maga2020* were also used frequently in this dataset.

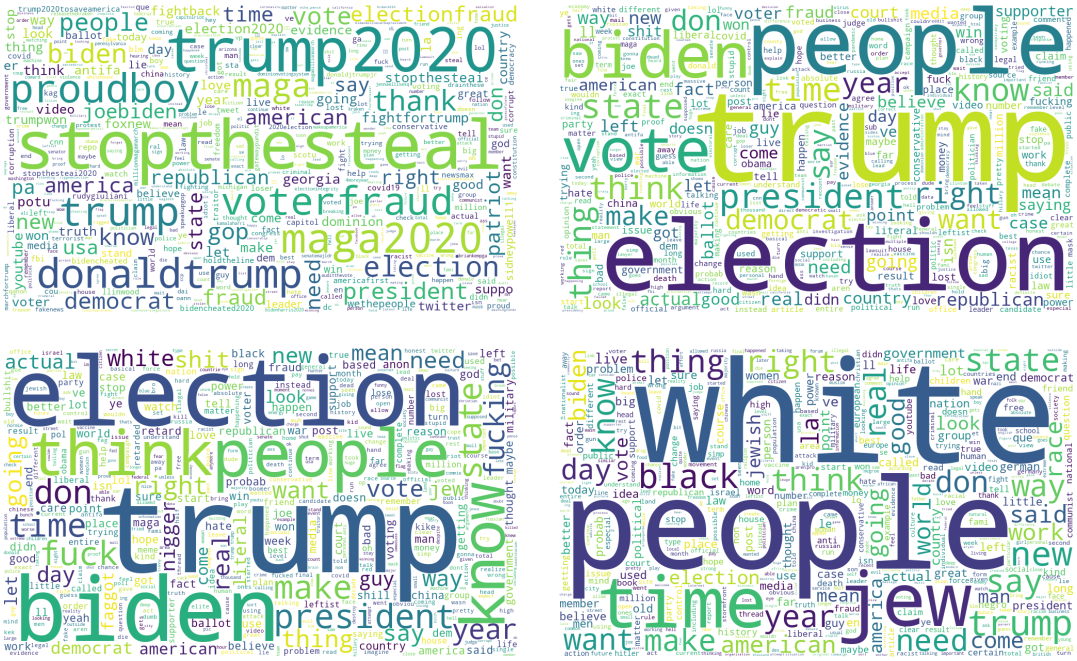

Figure 5.2: Word clouds of the most commonly used words across each election-related dataset: Twitter (top-left), Reddit (top-right), 4chan (bottom-left), and Stormfront (bottom-right).

A topic model, using the LDA topic detection model, also provides further insight into the most discussed subjects within each dataset, with the five identified

topics and the percentage of posts containing them being listed in Table 5.1; to get this percentage, the dominant topic out of the five topics was extracted in each post, and a cumulative total of the number of posts for each topic was calculated and represented as a percentage of the total posts in the dataset. When comparing the topics from all the datasets, it can be seen that similar topics are generally discussed or mentioned. For instance a common topic in the Twitter, Reddit and 4chan datasets is voter fraud. This topic is shown to be discussed extensively on Twitter (Topic#1, Topic#2, Topic#4) and Reddit (Topic#2, Topic#5). The Twitter posts also make frequent use of the slogan "stop the steal". Another topic that is common amongst the Twitter and Reddit datasets is the "march for Trump", later known as the capitol riots. The Twitter dataset in particular promotes details of how the march would be conducted, including the date, time and location, and also mentions 'excitement' for it (Topic#3). Both Twitter and Reddit also refer to the capitol riots as "republicans duty" and "election defense" respectively.

Similarly to what was shown in the word clouds, the 4chan posts evidently make use of more explicit and derogatory language than any of the other datasets of posts. The topic model shows that such terms seem to especially be used to discuss Trump losing the presidential election (Topic#2), suggesting such language is used more when voicing frustration. Out of all four datasets of posts, the Stormfront dataset is the only one that appears to discuss other topics outside of the election. Topic#3 in the topic model depicts users being "happy" and "proud" to "welcome new members" to the platform, where members are described as "patriots". This welcoming of new members could indicate the increase of users on this platform, though further analysis is required to confirm this. Similarly, Topic#5 in this dataset also exhibits key aspects of the identity and values of Stormfront users, including "nationalist", "white", and "segregation of races". Again, this topic shows the pride that Stormfront users have of this shared identity.

To further assess the findings from the topic modelling, the proportion of posts from each platform containing all five of the topics were then examined, which is also included in Table 5.1. Although it is very much clear that the posts collected

Table 5.1: A topic model of the most discussed topics during the 2020 US election and the percentage of posts containing them.

	Twitter	**Reddit**	**4chan**	**Stormfront**
Topic#1	audit michigan, rigged attorney, viral michigan, completed forensic, blocking disclosure, state blocking **(11%)**	donald trump, support, register, vote, state vote **(20%)**	MAGA, awoo, awoo, MAGA hat, MAGA forever, MAGA 2020 **(25%)**	people, trump, member, time, white, world, stormfront, biden, nationalist **(15%)**
Topic#2	antifa, stop the steal, proud boys, march for trump, january, white **(18%)**	make, report voter, trump campaign, fraud, voter fraud **(15%)**	still, supporter, vote, lost, lose, going, trump, f***, n****** **(21%)**	trump, said, like, white, would, people, state, want **(18%)**
Topic#3	march for trump, excited, join, washington, tomorrow, coming **(36%)**	trump 2020, MAGA, liber tears, MAGA 2020, breaks **(17%)**	vote, count, case, state, election fraud, votes, voting, election day **(23%)**	welcome, white, member, stormfront, patriot, proud, white new, happy**(27%)**
Topic#4	voter fraud, camera busted, caught, republicans duty, stop the steal **(15%)**	election defense, contact state, trump march, stop washington **(23%)**	lost, trump lost, lost election, lol, lost biden, white**(20%)**	trump, white, election vote, state, would, like, biden, jews**(19%)**
Topic#5	january 6th, ellipse, rsvp, white house, 7am, 6th doors, join january **(20%)**	vote, ballot, trump, president, election, voter, people, state **(25%)**	watch, president, video, MAGA, youtube, ballot, border, capitol, trump **(11%)**	stormfront, white nation, segregation, proud, nationalist, black race **(21%)**

from all four platforms are Trump-supporting, posts from Reddit are shown to repeatedly encourage their audience to register and cast their votes (Topic#1 and Topic#5), with such topics being the most common within the dataset (appearing in around 40% of the posts). This could be due to the fact that these posts were collected from a subreddit (r/donaldtrump) that was obviously created to support Trump, but each of the other three platforms seem to discuss Trump and exhibit their support more in the aftermath of the election, through topics on voter fraud and "marching for Trump".

Adding to this, it can also be observed that the most prominent topics appearing within the Twitter dataset were those discussing the Capitol riots, where 56% of all the posts contained these topics (Topic#3 and Topic#5) — notably more than the other datasets — and a further 26% of the posts containing topics related to voter fraud (Topic#1 and Topic#4). This suggests that Twitter may have been used for encouraging participation and the coordination of the Capitol riots more than any of the other platforms, which could be due to the larger audience size that Twitter offers; albeit the heavier moderation. These findings are supported by the frequency of posting shown in Figure 5.1, where it shows that Twitter users seemed to post more during the time of the Capitol riots in the first week of January, in comparison to Reddit users.

5.3.3 Linguistic and Sentiment Analysis

In addition to identifying the main topics of discussion on each platform, this analysis was also interested in examining and comparing the sentiment and the linguistic composition of the posts. To gain insight into this, the LIWC linguistic analysis tool was used to highlight any key differences between each dataset, the findings from which aim to address RA3. The results from this analysis are summarised in Figure 5.3, Figure 5.4, and Table 5.2, where the mean percentage of all words within each set of posts that fall into a particular LIWC category is shown. Example words of each category can be found in [261].

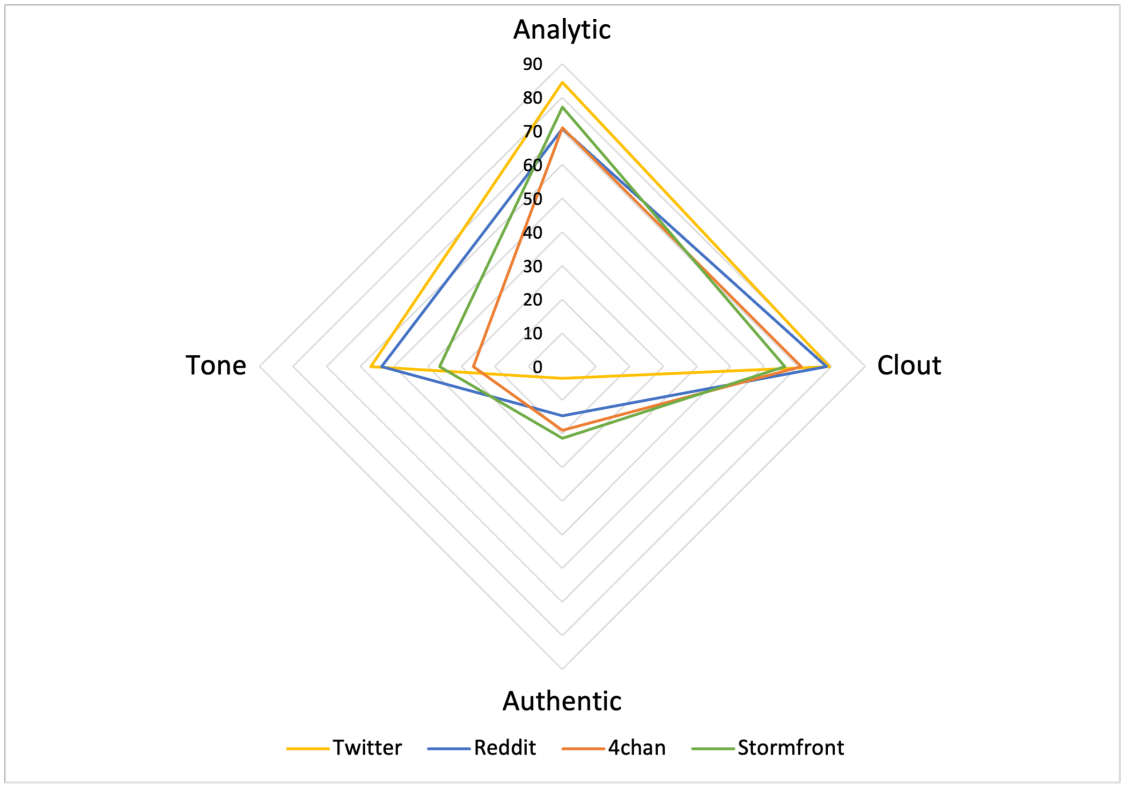

Figure 5.3: A comparison of the summary language LIWC categories across all four election-related datasets.

In the first stage of the LIWC analysis, the four summary language variables (*analytical thinking*, *clout*, *authenticity*, and *emotional tone*) were compared across each dataset. The results for this are shown in Figure 5.3. Through this, it can be observed that the Twitter posts have a higher score for *analytical thinking* (μ = 84.43) as compared to the other platforms. The *analytical thinking* score reveals the extent of analytical, logical and consistent thinking, in contrast to more intuitive, narrative writing [261, 268]. This suggests that users on Twitter would post more consistent thoughts and opinions during the 2020 US election. The LIWC analysis also shows that the *clout* scores for Twitter (μ = 79.37) and Reddit (μ = 78.58) are much higher than those of 4chan (μ = 70.96) and Stormfront (μ = 66.11). A higher clout score demonstrates a sense of authority and confidence [261].

In contrast, the *authenticity* score for the Twitter dataset (μ = 3.52) is considerably lower than the other three datasets, where the Stormfront posts have the highest score (μ = 21.31) for this particular attribute. This suggests that the

posts on Stormfront are more honest, personal and disclosing [268]. The scores for the emotional tone from 4chan ($\mu = 26.48$) and Stormfront ($\mu = 36.42$) are shown to be lower than 50, indicating the presence of a more negative tone, with 4chan posts being the most negative. On the other hand, the Twitter ($\mu = 56.86$) and Reddit ($\mu = 53.76$) scores indicate a generally positive tone.

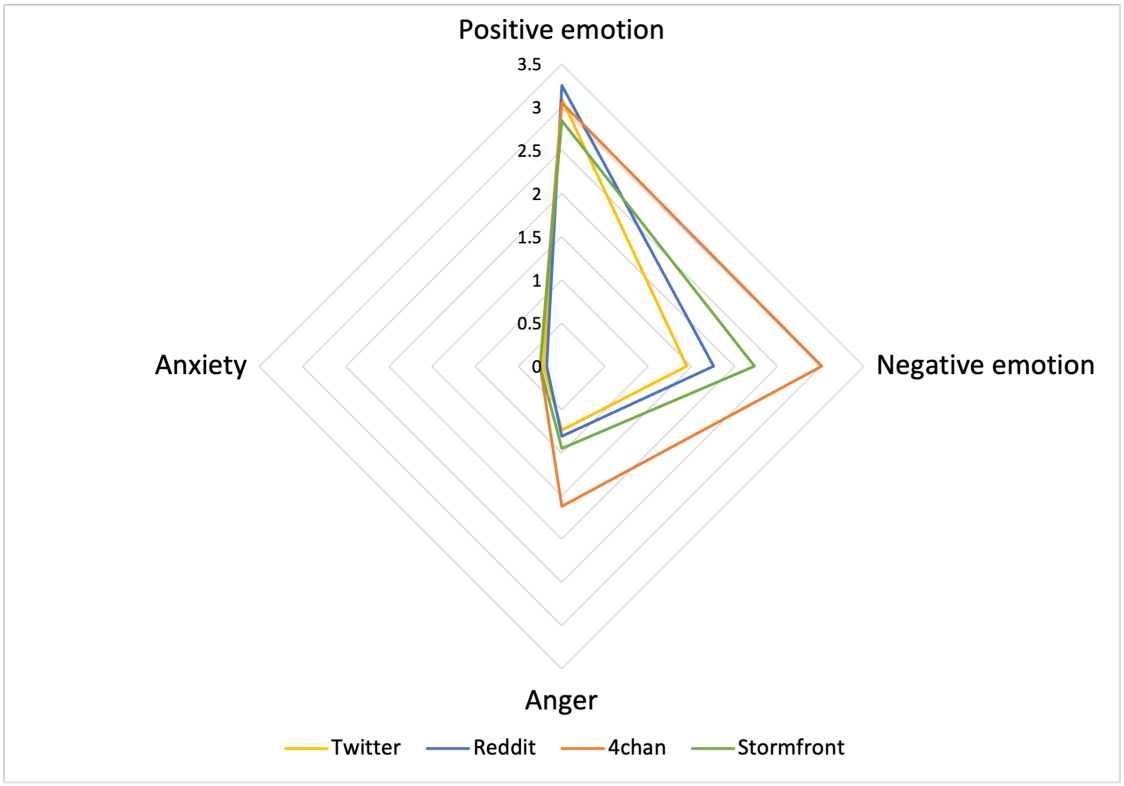

Figure 5.4: A comparison of the sentiment LIWC categories across all four election-related datasets.

The overall sentiments of each platform were explored further by using all the LIWC sentiment categories *positive emotion*, *negative emotion*, *anger* and *anxiety*, the scores for which are displayed in Figure 5.4. The results from this LIWC analysis show that all four platforms generally use more positive emotion than negative emotion, with posts from Reddit using positive emotion the most ($\mu = 3.25$). To gain further understanding of the context in which positive emotion has been used, a sample of posts were manually examined as well. This was able to show that Reddit posts encouraging users to vote and "support" Trump seemed to use more positive emotion. Since the topic modelling in Section 5.3.2 showed that these were

the most prominent topics identified in the Reddit dataset, it is clear how positive emotion is much more present than negative emotion on this platform.

Similarly, the results for the Twitter dataset also show that positive emotion (μ = 3.09) was used a considerable amount more than negative emotion (μ = 1.45). When further exploring the tweets using such sentiment, the study found that it was mostly used to rally people to participate in the Capitol riots, and their "excitement" and "happiness" that they were working to make a change — this was again shown to dominate much of the topics of discussion and is in line with previous findings from the topic modelling conducted. On the other hand, it is also clear to note from the LIWC analysis that negative emotion is used more in the 4chan posts (μ = 3.01) than any other dataset. From assessing a sample of the posts using negative sentiment, this work finds that this emotion was mainly used when expressing frustration over Trump's loss in the election. The topic modelling conducted in this chapter also showed that discussing Trump's loss was a prominent topic of discussion on 4chan. Additionally, similar results can be seen when looking at the level of *anger* in each dataset from the LIWC results, where 4chan posts make use of such language much more than any of the other platforms (μ = 1.62).

Table 5.2: Results from the linguistic analysis of the election-related datasets using LIWC.

LIWC Category	Twitter	Reddit	4chan	Stormfront
I	1.20	1.69	1.60	1.71
We	1.45	1.02	1.40	0.85
You	1.32	2.34	1.29	1.04
They	0.45	0.81	1.42	1.36
Swear	0.15	0.39	1.31	0.17
Religion	0.42	0.15	0.58	0.75
Money	0.62	1.25	0.76	0.77

The overall functional compositions of the posts in each dataset is then explored, with particular regard to the pronouns that have been used. The results for this analysis are detailed in Table 5.2. On the whole, most of the platforms would use more first-person singular pronouns (such as *I*, *me*, *my*) than first-person plural pronouns (such as *we*, *our*, *us*). The only exception to this is the Twitter dataset,

for which the opposite is true, as it used the least amount of first-person singular pronouns ($\mu = 1.20$) and the most amount of first-person plural pronouns ($\mu = 1.45$). Second-person pronouns (such as *you*, *yours*, *yourself*) were particularly more present within the Reddit posts ($\mu = 2.34$) than any other dataset. This is consistent with previous findings from this study, where the Reddit posts were frequently found to address their audience to encourage them to vote, for instance "make sure you go cast your vote".

When examining the usage of plural third-person pronouns (such as *they*, *them*), it is shown that they were used the most in the 4chan ($\mu = 1.42$) and Stormfront ($\mu = 1.36$) datasets, which indicates the presence of the "Us Vs. Them" dichotomy mentality described in [269]. This is consistent with the findings gained from the keyword and topic analysis carried out in the previous section, where 4chan and Stormfront posts in particular would frequently refer to groups of 'others', specifically Jewish people and black people, in degrading and discriminatory ways.

The use of pronouns has often been identified as a discursive tool used to persuade audiences in previous works. This is partly due to how different pronouns have a variable scope of reference, which is determined by the audience, who can then interpret whether they are inclusive or exclusive of them [270]. In particular, the use of personal pronouns such as *we*, *you*, *our* and *us* is a common persuasive technique to make audiences feel more immediately included. The LIWC analysis shows that this particular strategy is used more in the Twitter and Reddit posts than the other two platforms, suggesting that there is more of a sense of community on these platforms. Similar observations were made in previous studies exploring the usage of pronouns in hate speech on Twitter [271]. In this regard, the Stormfront posts are thus the least inclusive of their target audience.

These inferences are also corroborated by the level of *clout* within each dataset, where the Twitter and Reddit datasets had the higher scores. This suggests that the posts on these two platforms are composed in a way that would be more influential to their target audiences. Again, this is also evident in the previous findings of this

study as both Twitter and Reddit try to encourage their audiences to take specific actions, including participating in the Capitol riots and casting their votes.

Other notable observations from the LIWC analysis include the usage of *swear* words, which was used the most in 4chan ($\mu = 1.31$), confirming the observations made from the previous analysis. Another noteworthy linguistic component is the mention of *money*. In this case, Reddit posts were shown to use more money-related language than the other datasets ($\mu = 1.25$). It was later found during the URL domain analysis, some of the findings for which are detailed in Chapter 6, that Reddit posts would frequently post links to fundraising initiatives for Trump's campaign, and would encourage other users to donate.

5.4 Case Study 2: COVID-19 Pandemic

5.4.1 Participation Trends

To address RA1, the analysis framework first examines the frequency of content posted during the COVID-19 pandemic across the collection period in each of the four datasets. Figure 5.5 illustrates that, again, Twitter has significantly more content than the other platforms, with 4chan also having a considerably high volume of posts related to the pandemic. In contrast, it is clear that content is posted much less frequently on Stormfront. Within the context of COVID-19-related content, the r/donaldtrump subreddit had significantly fewer posts as compared to the amount of content posted to this subreddit during the 2020 US election.

To show how the amount of participation compares across all four platforms over the course of the collection time frame, standard score normalisation [264] is used to create a graph of all the normalised data from each dataset. In this case study, posting frequency is measured on a monthly basis to assess how real-time developments in the pandemic affected online discourse. For the most part, similar trends can be seen in the amount of participation on each platform over the collection period during COVID-19. The most steep peak can be seen on all four platforms around March 2020. This marked when COVID-19 was first declared as

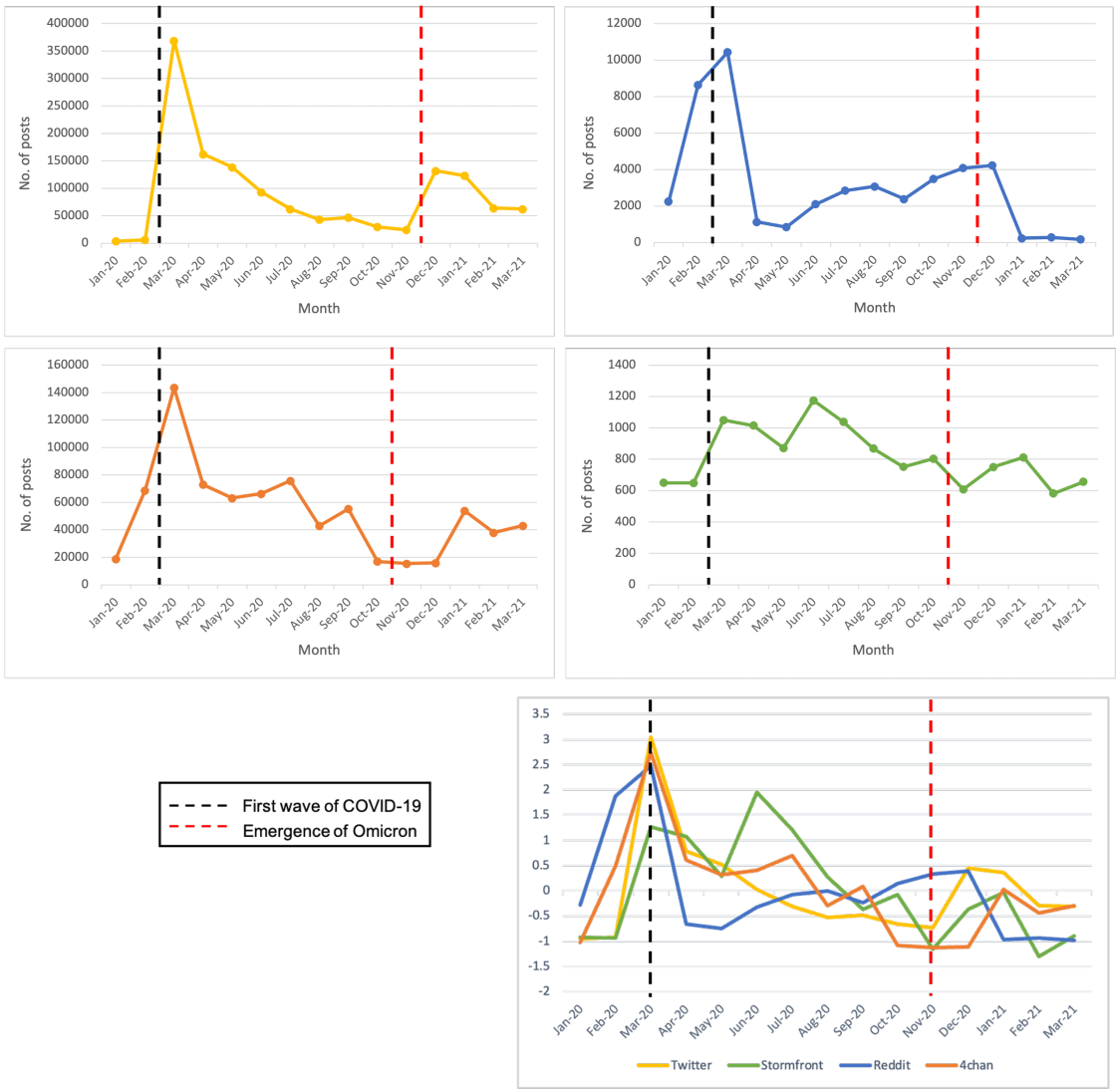

Figure 5.5: Graphs showing the frequency of posts across each dataset over the course of the COVID-19 pandemic: Twitter (top-left), Reddit (top-right), 4chan (bottom-left), and Stormfront (bottom-right). A graph with the normalised data from all four datasets is shown at the bottom.

a pandemic [253], as well as the introduction of severe restrictions and regulations, such as lockdowns and travel bans, throughout much of the world.

The Twitter, Reddit and 4chan datasets show a significant drop in posting activity shortly after this period, but also start to peak again around November 2020 and January 2021. This could indicate the time when lockdowns were once again enforced across many parts of the world, particularly during the Thanksgiving and Christmas holiday season, or when the Omicron variant of COVID-19 emerged, causing infections and death rates to reach unprecedented levels. These peaks

could also be attributed to the 2020 US election held in November, where the COVID-19 response and policies continued to be an important subject in political debates and campaign agendas.

As previously mentioned, the r/donaldtrump subreddit was banned due to its involvement during the January 2021 Capitol riots. As a result, there is a sudden halt in Reddit posts during the final months of the data collection period. Although there is a peak in the frequency of posts during the initial stages of the pandemic, posting behaviour on Stormfront is, again, mostly steady. This suggests that significant offline events do not have a considerable influence on the online behaviour of users on this particular platform. Instead users remain mostly consistent with their posting behaviour.

Therefore, in answer to RA1, the participation trends across all four platforms seem to be similar, in that they generally peak in posting frequency at around the same time periods following key real-time developments over the course of the COVID-19 pandemic. However, such events seem to control the discourse on platforms with larger audience sizes like Twitter, Reddit and 4chan, more than it does on smaller, underground platforms like Stormfront, where posting frequency is generally more consistent despite offline events.

5.4.2 Keywords and Topic Analysis

The cross-platform analysis framework is then applied to the COVID-19-related datasets to identify the frequently used words and topics on each platform, which are subsequently used to address RA2 by determining the dominant topics in the online discourse on each platform. The word clouds in Figure 5.6 demonstrate the most frequently used terms on the analysed platforms, with “China”, “Chinese”, “virus” and “COVID” being mentioned most often. Notably, “China” is mentioned much more than the other words on Reddit and 4chan in particular. This is likely due to the terms used to filter posts during data collection, though, again, this filtering process was used with all four datasets to ensure only content relevant to COVID-19 was collected.

Figure 5.6: Word clouds of the most commonly used words across each COVID-19-related dataset: Twitter (top-left), Reddit (top-right), 4chan (bottom-left), and Stormfront (bottom-right).

Given that the COVID-19 related posts on Reddit were collected from a Trump-supporting subreddit (r/donaldtrump), "Trump" is unsurprisingly one of the most mentioned words in the Reddit word cloud. "Trump" appeared more often on Reddit than on the other platforms, although he is also discussed frequently in the Stormfront posts. Similar to the findings from the analysis of Case Study 1, offensive and discriminatory language can be found in 4chan posts, particularly against black and Asian people. The word cloud for this dataset shows that derogatory language and profanity were used more frequently on this platform than on any of the other platforms.

Interestingly, the Stormfront dataset shows that terms such as "COVID" and "virus" were used notably less than terms related to white nationalism and anti-semitism, such as "white", "Jew", "black" and "race". This finding is, however, still consistent with previous literature exploring Stormfront, which has shown this platform to focus on promoting white-nationalist and antisemitic views. This keyword analysis also shows that Twitter users frequently used slogans linked to conspiracy theories and hateful movements, including referring to the pandemic

as the “plandemic”, “Chinese virus”, or “Wuhan flu”. The word cloud for this dataset additionally demonstrates an emphasis on blaming China and Asian people for the pandemic, with slogans such as “make China pay” and “china lied people died” being used frequently.

To gain further insight into the key subjects of discussion within each dataset, a topic model, using the LDA topic-detection model, is applied to all four datasets. The five identified topics and the percentage of posts containing them are listed in Table 5.3; this percentage is, again, derived by extracting the dominant topic in each post, and the total sum of the number of posts for each topic is then represented as a percentage of the total posts in the dataset. The topic model shows that the most common topic in COVID-19-related posts from all four platforms is directing a majority of the blame for the outbreak towards China. Users on each of the platforms often express this by frequently using the terms “Chinese virus” and “Wuhan virus” to refer to the pandemic.

Twitter, Reddit and 4chan also communicate particular frustration with China through aggressive terms. Such topics were discussed in around 51% of the Reddit dataset and around 41% of the 4chan dataset, where the Reddit posts often include phrases like “f*** China” in COVID-19 related discussions (Topic #5), while 4chan posts would even go as far as suggesting to “nuke China” (Topic #5). As shown with the word cloud, Twitter users would also use slogans like “make China pay” frequently along with discussing how China has been unpunished, so as to demand that China be held accountable for the pandemic (Topic #4). This is consistent with the findings gained from the cross-platform analysis in Case Study 1, which showed how Twitter was often used to petition for various social movements and calls for action. In addition to this, conspiracy theories blaming China for the origin of the virus and for silencing “whistleblowers” speaking out against China’s role in the pandemic are also prevalent on Twitter (Topic #2).

The topic model for the Reddit and 4chan posts show that, overall, similar topics were discussed on both platforms. For instance, resistance to wearing face masks is a common theme on Reddit (Topic #3) and 4chan (Topic #1), along with

Table 5.3: A topic model of the most discussed topics during the COVID-19 pandemic and the percentage of posts containing them.

	Twitter	**Reddit**	**4chan**	**Stormfront**
Topic#1	chinese virus, uk coronavirus, stay home new york, national response, failed national, china, lockdown extension **(19%)**	china, china virus, china as*****, russia, communist, china flu **(11%)**	mask, wear, wearing mask, don't wear, f*** masks, face mask **(18%)**	jew, like jews, media, trump, jew york, jew owned, owned media **(21%)**
Topic#2	china, virus, world outbreak, help contain, refused help, whistleblowers china, silenced whistleblowers, virus originate **(23%)**	trump, people, like, just think, Biden, won't **(26%)**	virus, corona virus, fake, chinese virus, corona spread, flue, vaccine, wuhan **(25%)**	people, white people, white hate, blacks, black people, think, racist **(23%)**
Topic#3	agenda21, plandemic, reclassified, recovery rate, flawed, deaths, reclassified **(21%)**	mask, wear, wearing mask, don't wear, face cover **(19%)**	covid 19, deaths, died, coronavirus, cases, flu, vaccine **(15%)**	covid, covid19, vaccine, news, coronavirus, positive, trump, world, covid vaccine **(17%)**
Topic#4	outbreak, enormity, covering severity, communist party, china unpunished, criminal drtedros, make china pay **(22%)**	covid, covid19, deaths, died, covid deaths **(23%)**	asian, white, asian women, black, white men, ch*** **(19%)**	china virus, coronavirus, jews, flu, chinese virus, trump, ccp virus **(19%)**
Topic#5	document revelations, significant document, chinese virus, crisis, wuhan coronavirus **(15%)**	f*** china, china flu, wihan virus, f*** chinese, spread, deadly, coronavirus **(21%)**	chinese virus, jews, communist, ccp, chinese government, f*** china, nuke china, war **(23%)**	anti white, white hate, racist, racism, media, racist media. **(20%)**

discussions about the death rates of the pandemic (Reddit Topic #4, 4chan Topic #3). However it should be noted that both the word clouds in Figure 5.6 and the topic model display how 4chan frequently uses discriminatory and hateful language, particularly when degrading or complaining about various minority groups. This is in accordance with the findings from analysing Case Study 1, where the lack of moderation along with the culture of trolling and shock value on 4chan has resulted in the regular usage of hate speech and discriminatory language.

Additionally, within the context of the COVID-19 pandemic, previous studies have found that 4chan has been a hub for the spread of misinformation using derogatory terms and racist stereotypes [266]. Previous findings from Cinelli et al. also demonstrate how users skewed towards more questionable content or misinformation were more prone to using inappropriate, violent or hateful language [225]. This includes numerous posts that blame Chinese people or Asians in general for the spread of the virus, as well as promoting the narrative that certain minority groups are immune to the virus or are somehow responsible for spreading it intentionally [272].

Trump is, of course, a significant topic of discussion on Reddit (Topic #2), and he is also mentioned in most topics of discussion on Stormfront (Topic #1, Topic #3, Topic #4), with 57% of the posts in this dataset discussing topics which mention Trump in some way. Trump was a prominent figure in the public discourse surrounding the pandemic, and his statements and actions often drew criticism and controversy [273]. His use of divisive language and rhetoric during the pandemic, more specifically, referring to COVID-19 as the "Chinese virus", "Wuhan virus" and "Kung flu", has been shown to embolden and amplify extremist and hateful views surrounding the pandemic [274]. To further explore the context in which Trump was mentioned on Stormfront, a sample of the Stormfront posts were manually examined. Through this analysis, it became apparent that Trump is often viewed as the "only hope" in politics for white nationalists.

Similar to the findings from Case Study 1, the majority of topics identified in the Stormfront dataset are centred around pro-white and neo-Nazi ideologies,

where a strong sense of white identity and membership to the online community is evident amongst the users; around 41% of Stormfront posts contain topics related to such themes. Many posts feature antisemitic narratives and conspiracies, including the idea that white people are under threat from non-white groups, and that this threat is being orchestrated by a Jewish conspiracy [275]. This can be seen in Topic #1 and Topic #4 in the topic model, which promote the belief that Jews control the media, government and financial institutions, and are using their power to undermine the white race. Additionally, users often victimise themselves and advocate against white hate while referring to groups of 'others', particularly "blacks" and "Jews", as the enemy (Topic #2).

Within the context of COVID-19-related posts, the topic model shows that Twitter hosts several conspiracy theories, highlighting the impact of misinformation on discourse surrounding the pandemic (Topic #2, Topic #3, Topic #4). This is especially evident through frequent usage of the terms "agenda 21" and "plandemic". Agenda 21 is a non-binding UN resolution providing a comprehensive plan of action to be taken by governments and major groups regarding human impacts on the environment. Though this resolution aims to promote environmentally friendly practices, conspiracy theorists linking agenda 21 to COVID-19 became prevalent during the pandemic. These often suggest that COVID-19 is being used as a pretext to enforce government control over citizens through the various restrictions and regulations [276]. Similarly, the term "plandemic" has often been used by conspiracy theorists to refer to the false and baseless claim that the COVID-19 pandemic was not a naturally occurring event, but rather a planned and intentional event by certain individuals or organisations [277]. Such narratives have contributed to the spread of hateful and discriminatory attitudes and actions towards certain groups, especially minority groups. Over 60% of the Twitter posts would discuss topics related to false narratives and conspiracies.

5.4.3 Linguistic and Sentiment Analysis

The next component of the cross-platform analysis in Case Study 2 comprises of exploring and comparing the sentiment and linguistic composition of the COVID-19-related posts collected from each of the four platforms. Again, LIWC is used to highlight the key differences in the psychological processes and various linguistic dimensions, the findings from which are used to address RA3. The results from this analysis are shown in Figure 5.7, Figure 5.8, and Table 5.4, which include the mean percentages of all words within each set of posts that fall into a particular LIWC category. Further details on how these categories are calculated as well as example words can be found in [261].

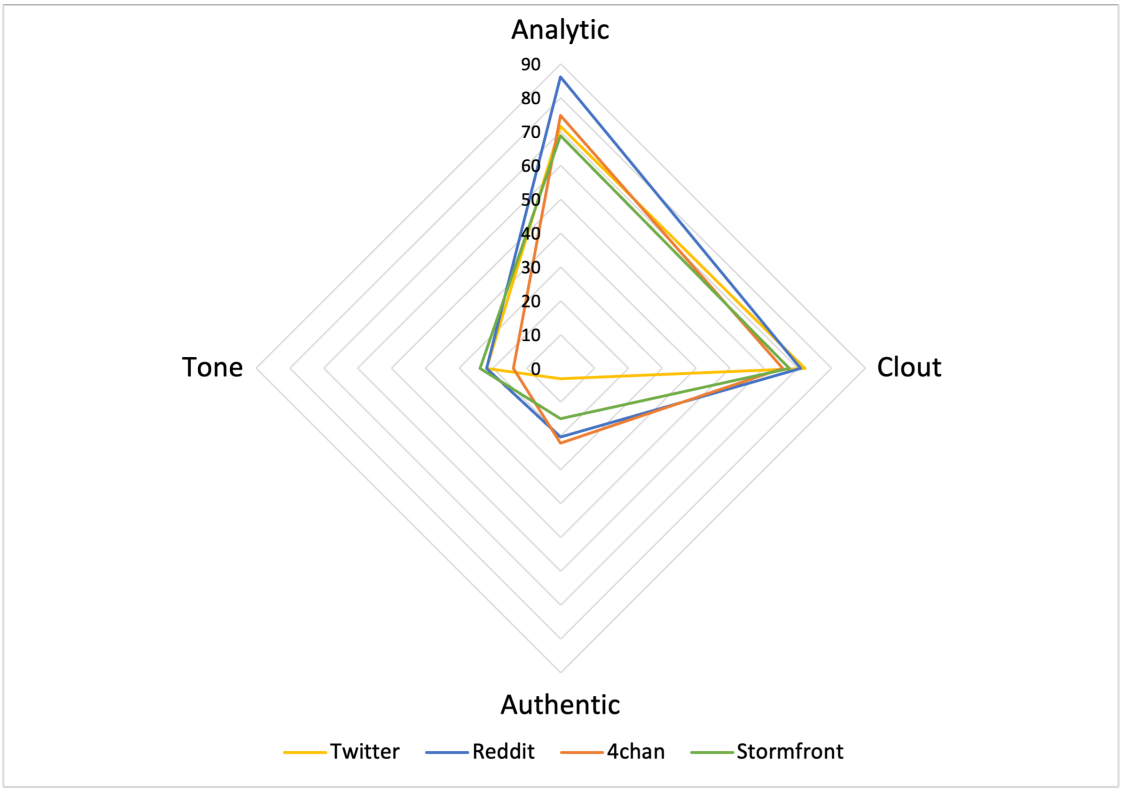

Figure 5.7: A comparison of the summary language LIWC categories across all four COVID-19-related datasets.

Firstly, the four summary language categories (*analytical thinking*, *clout*, *authenticity*, and *emotional tone*) were compared across each dataset. The results for this are shown in Figure 5.7. The LIWC analysis with the COVID-19-related posts show that the Reddit dataset has a much higher score for *analytical thinking* than the other

platforms ($\mu = 86.16$), indicating that users would post more consistent thoughts and opinions on this platform. In comparison to Case Study 1, where the Twitter dataset had the highest *analytical thinking* score ($\mu = 84.43$), Twitter posts had a much lower score in this category ($\mu = 74.72$), suggesting a lower degree in logical and hierarchical thinking. This could partly be due to the fact that, over the course of the pandemic, Twitter posts were shown to host a number of false narratives and conspiracy theories, as shown in the findings from the topic model in the previous section. Similar to the results from Case Study 1, the degree of *clout* within the Twitter ($\mu = 72.16$) and Reddit ($\mu = 70.96$) datasets are shown to be higher than 4chan ($\mu = 65.88$) and Stormfront ($\mu = 67.58$). As mentioned previously, the higher *clout* scores demonstrate a stronger sense of authority and confidence [261].

Again, the *authenticity* score for the Twitter dataset ($\mu = 2.98$) is much lower than the scores for the other platforms. Interestingly, the Stormfront dataset also has a lower score for *authenticity* ($\mu = 14.92$) when compared to Case Study 1. The Reddit ($\mu = 20.34$) and 4chan ($\mu = 22.04$) posts, however, have considerably higher scores in this category. One inference that could be made from this is that the posts on these two platforms are more personable and disclosing [268]. The scores for the *emotional tone* of each dataset are all below 50. This indicates that the overall emotions on all four platforms are negative. This is consistent with the findings from the keyword and topic analysis detailed in the previous section, where the majority of users discussed frustration with the various actors, particularly China, they blamed for the cause of the pandemic. Similar to Case Study 1, the 4chan posts were shown to be the most negative ($\mu = 13.90$).

The LIWC sentiment categories (*positive emotion*, *negative emotion*, *anger* and *anxiety*) were then used to further explore the sentiments of the posts from each platform, the results for which can be shown in Figure 5.8. Notably, the findings from this component of the LIWC analysis show that *negative emotion* was generally used a lot more than positive emotion across all four platforms, within the context of COVID-19. As expected, 4chan posts have the highest score for *negative emotion* ($\mu = 3.18$). To gain further contextual understanding on the

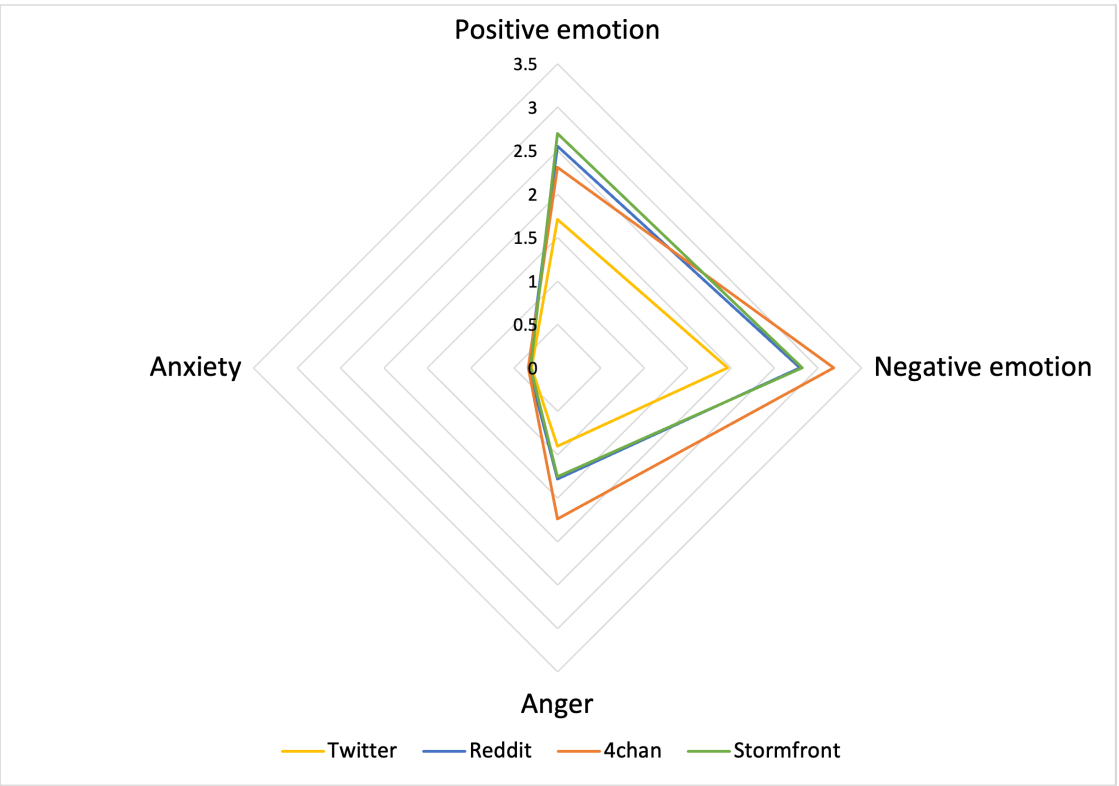

Figure 5.8: A comparison of the sentiment LIWC categories across all four COVID-19-related datasets.

usage of negative emotion, a small sample of posts were also manually analysed. Through this, it was shown that 4chan and Reddit posts complaining about the various regulations put in place to try to control the spread of COVID-19, such as wearing face masks, used more negative emotion. Stormfront users would use negative emotion when promoting hateful narratives against Jews controlling the media, and the victimisation they feel as targets of "white hate". As these were shown to dominate most of the topics of discussion in the topic modelling carried out in the previous section, it is unsurprising that negative emotion was much more present than positive emotion in COVID-19-related posts from all four platforms.

From this analysis of the general sentiment of the posts on each platform, it can be observed that the *anxiety* scores for all four datasets are very similar. However, when these scores are compared to the findings from the LIWC analysis in Case Study 1, it is clear that *anxiety* levels are overall higher across all the datasets. This is most likely due to the discussion of impacts of the pandemic, such as increasing death

rates as well as the fear-mongering caused by various conspiracy theories, which were shown to be prevalent themes in the topic analysis discussed in the previous section.

Similar to the results from the analysis of *negative emotion*, the scores for the *anger* LIWC sentiment category show that the 4chan dataset had the highest level of *anger* ($\mu = 1.74$). The scores for this category were also generally higher in the COVID-19 datasets from all four platforms than the election datasets in Case Study 1. Again, such language was used to express frustration with regulations put in place, and to put the blame of the creation and spread of the virus on various groups, namely China and Asians, through conspiracy theories and false narratives. 4chan posts were particularly aggressive in expressing this as they would often discuss how to retaliate against the virus by "nuking China", which was previously shown to be a major topic of discussion on this platform in Section 5.4.2.

Table 5.4: Results from the linguistic analysis of the COVID-19-related datasets using LIWC.

LIWC Category	Twitter	Reddit	4chan	Stormfront
I	0.52	1.64	1.57	2.03
We	0.88	1.00	0.87	0.94
You	0.71	1.73	0.86	1.59
They	0.54	1.42	1.68	1.67
Swear	0.17	0.54	1.29	0.25
Religion	0.28	0.21	0.39	1.30

The next stage of the LIWC analysis explored the functional composition and linguistic dimensions of the posts within each dataset. In particular, this analysis examined the usage of pronouns. The results from this analysis are included in Table 5.4. In general, most of the platforms would use more first-person singular pronouns (such as *I*, *me*, *my*) than first-person plural pronouns (such as *we*, *our*, *us*), as was the case in the analysis for Case Study 1. Again, in terms of the Twitter dataset, the opposite was true, where posts used less first-person singular pronouns ($\mu = 0.52$) than first-person plural pronouns ($\mu = 0.88$). In comparison to the LIWC analysis carried out in Case Study 1, second-person pronouns (such as *you*, *your*, *yourself*) were used considerably less in COVID-19-related posts, with Twitter posts using them the least ($\mu = 0.71$). This could largely be due to users not directly

addressing their audiences as much across all platforms within the context of the pandemic, unlike in Case Study 1, where users would often direct their audiences to vote, protest or donate to fundraising campaigns.

The LIWC analysis also shows that third-person plural pronouns (such as *they*, *them*) were, again, used the most in the 4chan ($\mu = 1.68$) and Stormfront ($\mu = 1.67$) datasets. This indicates a larger presence of the "Us vs. Them" mentality [269] on these two platforms than on Twitter and Reddit. This is consistent with the findings gained from the keyword and topic analysis carried out in the previous section, where 4chan and Stormfront posts would especially refer to groups of "others", specifically Chinese, Asian people and Jews, in discriminatory and hateful ways. When comparing the scores for third-person plural pronouns in Case Study 1 and Case Study 2, it is clear that a higher quantity of these pronouns can be found in the COVID-19-related posts in Case Study 2.

As noted earlier, previous works have often identified the use of pronouns as a discursive tool used to persuade audiences, partly due to how they can be interpreted by the audience on whether they are inclusive or exclusive of them [270]. In particular, the usage of personal pronouns (such as *we*, *you*, *our*, *us*) is a frequently utilised persuasive tactic that can help make an audience feel more included. The LIWC analysis reveals that this particular approach is used more in Reddit and Stormfront posts compared to the other two platforms, indicating a stronger sense of community on these sites. This finding aligns with previous observations made in the topic analysis, which found that Stormfront users exhibited a stronger sense of shared white identity and membership to their online community.

The Twitter posts, however, were shown to be the least inclusive of their audience, which could be attributed to the platform's larger audience size. The platform structure of Reddit and Stormfront as forums that foster online communities may also account for these differences observed in the use of personal pronouns. This is in contrast to the results of Case Study 1, where Twitter was found to use personal pronouns more often. This could be due to users primarily being focused on posting

updates about COVID-19 and promoting conspiracies and false narratives, rather than campaigning for specific actions.

Other notable observations from the LIWC analysis include the usage of *swear* words, which were used the most in 4chan ($\mu = 1.29$), similar to previous observations. This finding corroborates the observations made in the keyword and topic analysis. Another noteworthy dimension highlighted in the LIWC analysis is the mention of *religion*. In this case, Stormfront posts ($\mu = 1.30$) used considerably more religion-based language than the other datasets. This is consistent with previous findings given that the Stormfront posts were shown to mostly promote various antisemitic conspiracies and narratives. Previous studies have also established how Christianity plays a central role in the identity of Stormfront users [275]. The scores for *religion* were shown to be higher in Case Study 2 than in Case Study 1, which can largely be attributed to the fact that antisemitic narratives were much more prevalent in COVID-19-related discussions, specifically including the false narrative that Jews were behind the pandemic, either by creating the virus as a biological weapon or by using the pandemic to further increase their own interests [278].

5.5 Summary

In this chapter, the cross-platform analysis framework was used to investigate how hateful behaviours and content compare across different online platforms, so as to gain a clearer understanding of the dynamics of the global hate ecosystem. More specifically, this involved applying the analysis framework to data collected over two case studies, the 2020 US election and the COVID-19 pandemic, from Twitter, Reddit, 4chan and Stormfront. This research builds on previous work by exploring both mainstream platforms, such as Twitter and Reddit, as well as non-moderated fringe platforms, like 4chan and Stormfront. The analysis framework harnessed various computational methods, including topic modelling, linguistic analysis and sentiment analysis to explore the type of content that is promoted on each platform.

The analysis for both case studies focused on three research aims: investigating the participation trends across the four platforms during the 2020 US election and

COVID-19; identifying and comparing the main topics of discussion on each of the platforms; and comparing the differences in the sentiment and overall linguistic composition of the posts in each dataset. The analysis revealed that the participation trends on all four platforms are generally very similar, with peaks occurring at corresponding times to real-time developments. However, such events seem to control the discourse more on platforms with a larger user base, like Twitter, Reddit and 4chan, as compared to smaller, underground platforms like Stormfront. This was demonstrated in both case studies, where Stormfront had the most consistent posting behaviour despite offline events.

Through topic modelling, this analysis was able to find that Twitter and Reddit users harnessed the bigger audience size offered by the platforms to coordinate the organisation of the Capital riots, and encourage their audience to cast their votes for Trump. In Case Study 2, all four platforms would refer to the COVID-19 pandemic as "Chinese virus" or "Wuhan virus", which previous articles have linked to racist and hateful narratives [48]. In this context, Twitter and Stormfront were shown to predominantly promote false and hateful conspiracy theories, whereas 4chan and Reddit users mostly expressed their frustration with regulation and restrictions related to the pandemic. Finally, further sentiment and linguistic analysis showed the use of personal pronouns, which previous literature have shown to be a common persuasive technique in writing, were harnessed by platforms like Twitter and Reddit during the election to encourage their audience to vote for Trump or join the Capitol riots, as well as by Stormfront users over the course of the pandemic to promote a stronger sense of community and shared identity. The dichotomy mentality of "Us Vs. Them" is reflected strongly in 4chan and Stormfront, which often exhibit hateful narratives regarding groups of "others".

Overall, this chapter shows how factors such as platform affordances and contemporary social events occurring in real-time can affect the content and posting behaviours of hateful users on various platforms. These factors shape the discourse, coordination, and messaging strategies across the platforms. Informed with the results of this analysis, in the next chapter, the cross-platform analysis framework

is applied to the same two case studies to gain further insight into the content sharing and network dynamics of hateful groups and users on Twitter, Reddit, 4chan and Stormfront.

6

Identifying Networks of Hate Across Multiple Platforms

6.1 Introduction

Social-media platforms enable fast and widespread dissemination of information that can be exploited to spread hateful content. This popularisation of internet-based communication and interactivity of online platforms has further resulted in increased interest in understanding social interactions on a large scale. Several studies have explored various methods for identifying coordinated campaigns on a single platform, where they study how content spreads across a network of inter-connected users. As has already been demonstrated, online hate spans multiple platforms, thus there is a recognised need to jointly analyse the networks of users and shared content, as well as the diffusion of content across different sources, such as social networks and online forums. The aim here is, therefore, to develop methods for modelling and creating rich diffusion networks across multiple platforms.

In particular, the characterisation of the online networks, as well as the dynamics of information propagation in social-media platforms, blogs and other online forums has provided novel insight into how users can influence the behaviours of others over a variety of use cases, including marketing strategies and the study of disinformation. Mostly, these studies tend to focus on users from one particular platform, such as

Twitter. Several models have been introduced to identify influential users in a given online environment, and how they can impact content dissemination, including TwitterRank [279], ProfileRank [280] and Influence-Passivity [281], which rely on social-network structures and content relevance to detect influential nodes.

Though research in online hate has been fragmented across several disciplines, academic literature has only recently recognised that online hate is not simply an issue for a select few platforms, rather networks of hate are often linked across these platforms, forming a global 'network-of-networks' dynamic [11]. These networks formed by hate groups have proven to be remarkably resilient, and have increasingly shown to migrate across various platforms and other networks, maintaining and often expanding their connections in the process [11]. For instance, Zannettou et al. [29] demonstrate how various web communities can impact and influence each other by investigating how mainstream and alternative news propagate across multiple online communities. Using a statistical model, they highlight that small "fringe" online communities within Reddit and 4chan can have a substantial impact on large mainstream online communities like Twitter, as such online platforms are clearly not independent.

This chapter aims to further explore the network and content diffusion dynamics of online hate across Twitter, Reddit, 4chan and Stormfront within the context of the 2020 US election and the COVID-19 pandemic. More specifically, this chapter harnesses the cross-platform analysis framework by applying various network-analysis techniques to explore the type of content that is promoted on each platform, and to examine whether there is any cross-platform content sharing, namely in the form of common references to domains and subdomains. The findings from this analysis will then be used to gain insight into content diffusion across different social-media platforms and forums over the course of these two events.

The contributions of this chapter are as follows:

- NLP techniques are used to provide a comparison of the type of content that is most frequently shared on all four platforms through the usage of URLs.

- URL co-occurrence within every dataset is explored so as to understand the dynamics of content-sharing behaviours on all four platforms.
- Network-analysis techniques are harnessed to investigate if any common information sharing takes place across platforms, and thus provide a wider view of the content, particularly URL domains, that contribute to a larger ecosystem of hate. This is then used to gain insight into cross-platform content diffusion.

6.2 Methodology

This network and content diffusion analysis of hateful discourse over the 2020 US election and COVID-19 pandemic on the platforms Twitter, Reddit, 4chan and Stormfront is carried out with particular regard to following research aims:

- **RA1:** Compare the type of content that is most frequently shared on all four platforms during the US election and COVID-19 pandemic through the use of URLs.
- **RA2:** Analyse the network dynamics of content-sharing behaviours on all four platforms, specifically in terms of URL co-occurrence within posts.
- **RA3:** Determine if there is any evidence to confirm posting activity across platforms and explore the insights this provides into cross-platform content diffusion dynamics.

The details of the methods and analysis techniques used to address these research aims are discussed in the following subsections.

6.2.1 URL Analysis

The first research aim in this chapter (RA1) aims to identify and compare the type of content shared through URL domains on each of the four platforms. To address this research aim, every URL from each of the datasets collected is first retrieved so

as to extract the domains and subdomains from the posts on every platform. To verify whether any link shortener was utilised, each of the URLs were examined for sub-phrases from commonly-used link-shortening services (such as `bit.ly`, `ow.ly`, and `t.co` in the case of Twitter posts). Any shortened URLs were then expanded to obtain the full domain names using the Python 'urlExpander' package [282], where query terms at the end of each URL were also eliminated to acquire the base domain.

6.2.2 URL Co-Occurrence

RA2 aims to create network graphs of any URL co-occurrence that is found within each of the datasets, so as to compare content-sharing behaviours on all four platforms. Within the context of this analysis, URL co-occurrence refers to the presence of two or more URLs within the same post. The study of URL co-occurrence in SNA can help in understanding the structure and content of online networks, as well as the behaviour and interests of online users. The analysis of URL co-occurrence can also reveal the topical or thematic structure of online networks, as well as the behaviours and interests of online users. For instance, by identifying clusters of posts which express similar interests and preferences.

In order to create these network graphs of URL co-occurrence, the circle pack algorithm for visualising networks is used alongside the Louvain approach to community detection in online networks using the Gephi software package. The circle pack algorithm is a visualisation technique that can be used to represent large-scale network structures in a compact and intuitive way, and can be used in conjunction with the Louvain method to analyse community structure in large-scale networks [283]. This involves the identification of groups of nodes that are densely connected internally, but sparsely connected to nodes in other groups. The Louvain approach is one of the most widely used methods for community detection in SNA. This method is a modularity-based approach that optimises the assignment of nodes to communities [284]. The algorithm works by starting with each node in its own community, and then iteratively merging communities in a way that maximises the modularity of the network.

The circle pack algorithm is then used to visualise the community structure of networks detected by the Louvain method. The resulting visualisation shows each community as a separate circle, with nodes in each community arranged in a compact and intuitive way. The size of each circle represents the number of nodes in the community, and the distance between circles represents the strength of connections between communities. This visualisation aids in understanding the structure of large-scale networks, and identifies important communities and nodes within the network.

When analysing such URL co-occurrence networks, there are several measures or features that can provide valuable insights into the structure and function of the network. Below are some of the features that are considered within the analysis detailed in this chapter:

- *Modularity:* Modularity is a measure of the density of connections within communities compared to connections between communities. Higher modularity indicates a stronger community structure, and can be used to evaluate the effectiveness of the Louvain method in identifying distinct communities in the network.

- *Community size and composition:* The size and composition of each community can provide insights into the nature of the content being shared among the URLs. Are there a few dominant communities that contain the majority of the URLs, or are there many small, niche communities? The composition of each community can also reveal common themes or topics that are being discussed among the URLs.

- *Cluster Overlap:* The circle pack visualisation technique can reveal overlaps between clusters or communities, which may indicate common themes or topics that are being discussed across multiple communities. Overlaps can also reveal connections or similarities between seemingly unrelated communities.

Analysing these features of the URL co-occurrence network can help in identifying distinct communities and visualising their relationships.

6.2.3 Domain Network Analysis

To further examine the wider ecosystem of domain-sharing across all platforms, network-analysis techniques are utilised to construct "domain networks". This approach is adapted from Starbird's study exploring the ecosystem of alternate news domains on Twitter [285]. In this approach, domain networks are graph representations of the domains and subdomains shared within each platform and across other platforms. Here, two types of nodes are defined: one type of nodes (called platform nodes) represents each of the four platforms, and the other type of nodes (called domain nodes) represents the most frequently shared domains and subdomains in the datasets.

The edges in these domain networks represent where the domain/subdomains were shared. For instance, an edge between the domain node `youtube.com` and the platform node `Reddit` shows that this domain has been shared on Reddit. Domain nodes can have edges to multiple platform nodes, and vice versa. This allows inference of the coordinated effort by users from multiple platforms to amplify particular types of domains. Edge weight is determined by the frequency of the domains being shared on each platform. To trim the network, edges with weights less than five were removed to only include the most frequently shared domains and subdomains. This research makes use of the R package 'igraphs'[1] to create these domain networks, which were then used to gain insight into the content diffusion dynamics across multiple platforms.

6.3 Case Study 1: US Election

6.3.1 URL Analysis

To address RA1, the first aspect of this domain-network analysis involves exploring which domains were posted most frequently across all four election-related datasets, where the 10 most popular domain names have been listed in Table 6.1. After extracting all of the URLs included within each dataset of posts, it is immediately

[1] https://igraph.org/

apparent that the number of URLs being used within all four datasets is vastly different. The 4chan dataset included a total of 226,394 URLs, of which 40,256 were unique URLs. When calculating the ratio of posts that made use of a URL to share content, it was found that around 21% of 4chan posts would include a URL. In comparison, the Stormfront dataset used the least number of URLs, where only 4,057 URLs were used in total, with 2,279 of these being unique URLs. In this case, only 4% of the Stormfront posts were shown to include a URL.

In contrast, the Twitter and Reddit datasets made use of a considerably larger number of URLs. The Twitter dataset included a total of 1,479,936 URLs, where 292,561 were unique URLs, and around 98% of the Twitter posts including a URL. This analysis also found that, overall, URLs were used the most frequently within the Reddit dataset, with a total of 202,445 URLs being used, of which only 7,601 were unique, showing Reddit posts would consistently share links to the same sites. The ratio of posts in this dataset that made use of a URL also showed that 179% of the total posts included a URL, where most posts likely contained multiple URLs, or certain posts would mass-post several URLs. This is in part due to Reddit having a large character limit, so more content can be shared within posts than those from platforms such as Twitter, where the maximum character length is very limited.

Table 6.1: The most posted URL domains and their frequency across all four election-related datasets.

Twitter	**Reddit**	**4chan**	**Stormfront**
parler.com: 23,056	reddit.com: 34,786	pastebin.com: 7120	twitter.com: 156
TrumpMarch.com: 12,953	discord.gg: 24,973	promiseskept.com: 3601	jrbooksonline.com: 35
twitter.com: 11,461	armyfortrump.com: 9518	donaldjtrump.com: 3590	breitbart.com: 29
pscp.tv: 4895	vote.donaldtrump.com: 9515	magapill.com: 3588	gab.com: 27
theepochtimes.com: 3896	vote.gov: 9513	cbp.gov: 3371	jewworldorder.org: 23
youtube.com: 3782	shop.donaldjtrump.com: 9513	archive.is: 2963	bitchute.com: 17
electionevidence.com: 1032	trumpvictory.com: 9513	hereistheevidence.com: 2076	parler.com: 15
stopthesteal.us: 675	hereistheevidence.com: 7654	armyfortrump.com: 1861	davidduke.com: 13
gab.com: 672	pastebin.com: 6432	trumpvictory.com: 1856	nypost.com: 11
change.org: 525	archive.vn: 987	factba.se: 1630	thesun.co.uk: 8

By further examining which domains and subdomains were posted the most within each dataset, it was found that there are key distinctions in the types of domains shared. The Twitter dataset generally posted various news sources,

Trump-supporting domains, and content from other social-media platforms. News sources made up around 38% of the total URLs shared, where most of these news sources were right-leaning such as `foxnews.com` and `theepochtimes.com`, reporting on claims of voter fraud in the aftermath of the election. The Trump-supporting domains shared by the Twitter posts would also promote domains such as `TrumpMarch.com`, `stopthesteal.us` and `electionevidence.com`, which made up around 29% of the total URLs. Sharing links to content from other platforms made up around 16% of the total URLs, including `parler.com` and `gab.com`.

The posts from the Reddit dataset mostly promoted Trump-supporting domains (47%) like `trumpvictory.com` and `magapill.com`, which celebrated his achievements. These posts would also include fundraising initiatives for Trump's campaign (15%) like `shop.donaldjtrump.com`, as well as voter information and redirection (18%), such as `vote.gov`. Considering that the Reddit dataset included the most URLs but used the least number of unique URLs, this study may posit that Reddit users were much more proactive in their efforts to support Trump.

The posts within the 4chan dataset also mostly linked promotion-focussed domains (40%), such as the Trump-supporting domains used in the Reddit dataset. Around 17% of the total links were to archival repositories like `pastebin.com` and `archive.is`. Previous literature has established how archival sites have been used to save and preserve content that users may not want to be deleted or censored by mainstream social-media platforms [286]. This includes content that contains hate speech, disinformation, or other controversial material. By saving this content on an archival site, they can ensure that it remains accessible to others on the platform even if it is removed from mainstream platforms.

In contrast, the Stormfront posts seemed to share links to literature and articles on white-supremacist ideology (around 60%), including `jrbooksonline.com` and `jewworldorder.org`. Like the 4chan posts, they would also frequently post links to more explicitly hateful content on underground and uncensored repository sites, including `banned.video` and `archive.is`, as well as `.onion` sites. This is a strategy often used by far-right groups to promote their ideas and ideologies

outside of mainstream social-media platforms through more explicitly extremist or hateful material [206]. By using archival sites to redirect other users, they can circumvent censorship and reach a wider audience.

6.3.2 Analysing the Presence of URL Co-Occurrence

To address RA2, the cross-platform analysis framework is then used to further analyse the network dynamics of content-sharing practices on all four platforms. More specifically, this involved creating and examining network graphs of URL co-occurrence behaviours using the Circle Pack visualisation algorithm, as well as the Louvain approach to community detection. Figure 6.1 displays these URL co-occurrence networks for each dataset of election-related posts, where nodes represent the URLs being shared within posts, and edges represent when two or more URLs are are shared within the same post. Here, node size represents the number of times a URL has been shared within posts, where the larger the node, the more posts that have shared that particular URL. Any communities identified in this network analysis have also been highlighted and labelled in these network graphs.

The analysis of URL co-occurrence in each election-related dataset provided various insights into the content-sharing practices on all four platforms. Twitter was shown to have the most number of nodes within the network, which is unsurprising given URLs were shown to be a key component of content sharing in Twitter posts in Section 6.3.1, where this dataset had the highest number of URLs. Interestingly, very little URL co-occurrence can be seen in the network graph. This is most likely due to the small character-limit provided by Twitter, which would restrict users from sharing multiple URLs within the same post.

From applying the Louvain community-detection algorithm, the network graph also shows that there are around four major communities of URL co-occurrence, with certain nodes within these communities being much larger than others, indicating that certain URLs are shared by many posts alongside some other URL. Overall, the majority of the URL co-occurrence revolved around the ‘Stop the Steal’ or ‘March for Trump’ narrative. This includes a community of posts campaigning for

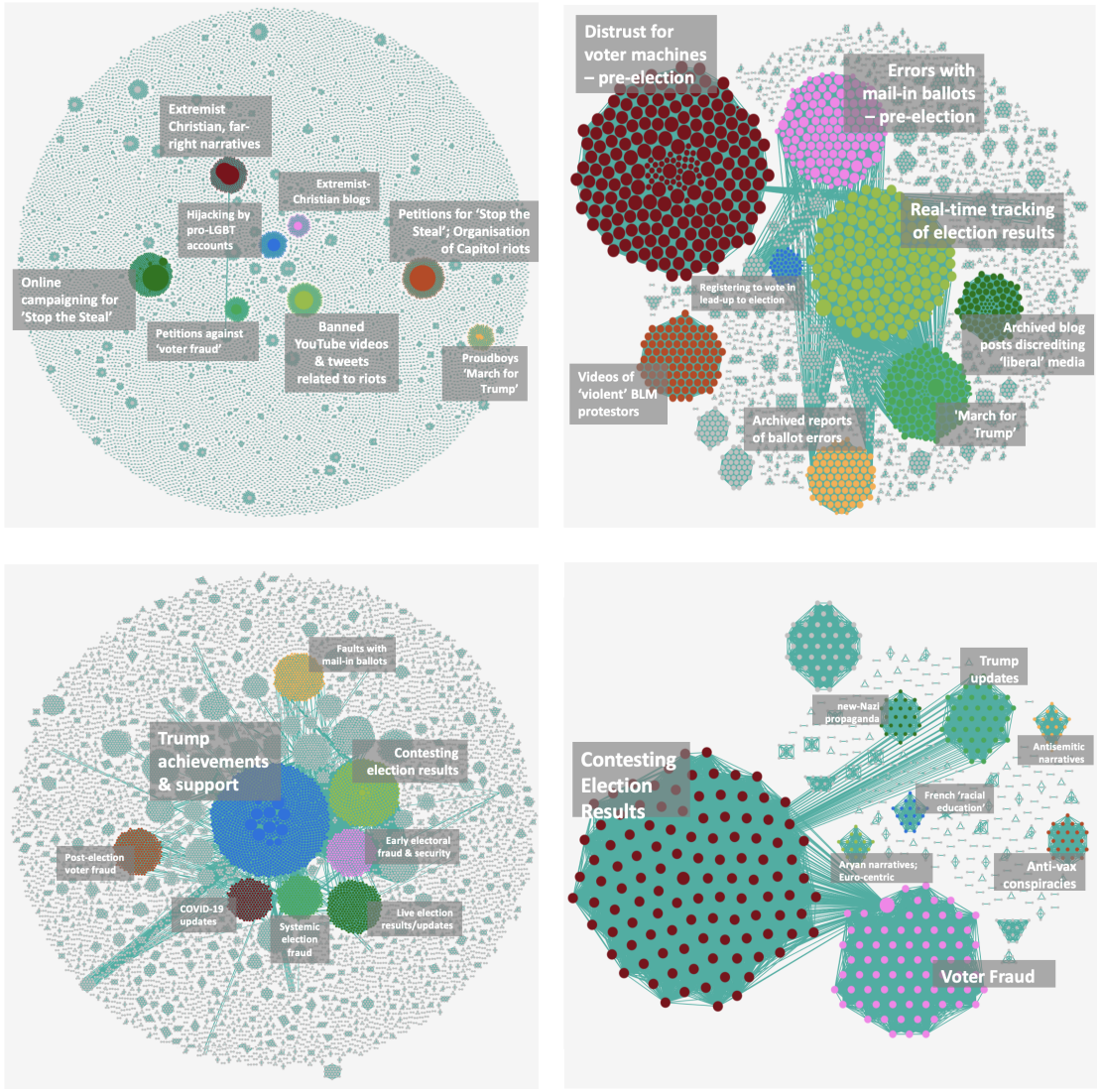

Figure 6.1: Network graphs of the URL co-occurrence behaviours found within each election-related dataset: Twitter (top-left), Reddit (top-right), 4chan (bottom-left), and Stormfront (bottom-right).

the 'Stop the Steal' movement, which was a social movement promoted by Trump and his supporters to protest that the election was allegedly stolen by Biden and the Democratic party [287]. The most shared URL within this community was to an account on the social-media platform Parler called "#StopTheSteal Caravan", belonging to a pro-Trump group that organised a rally (or 'caravan') of Trump supporters travelling to Washington D.C. for larger protests to contest the election results [288]. Other communities highlighted in the URL co-occurrence network show that online petitions to support the 'Stop the Steal' movement were also

circulated frequently amongst Twitter users.

In addition to these communities, a further notable observation that can be made from the URL co-occurrence network for the Twitter dataset is that one of the communities highlighted in the network graph includes a significant number of posts sharing links to a t-shirt company supporting the LGBT community (`lgbttees.myteespring.co`). Upon further inspection, it was noted that several accounts and hashtags associated with hate groups, such as '#proudboys', were hijacked by online activists during the height of the 'March for Trump' campaign, so as to take over hateful narratives by promoting pro-LGBT content instead. This has been a frequent occurrence in the past few years, as online users, mainly on mainstream platforms like Twitter, have realised that they can break up hateful discourse online by taking over hashtags predominantly used to promote hate [289].

The network graph for the Reddit dataset shows much more URL co-occurrence, with high modularity between the communities. Here, most communities of co-occurring URLs are based on distrust around the voting procedures in the lead up to the elections, including reports on supposed issues with mail-in-ballots. The co-occurrence network suggests that Reddit users would often post links to several of these reports at a time, given that there are so many edges connecting the nodes in the larger communities together. Another major theme in the content-sharing that can be observed on Reddit is registering to vote prior to the elections, through domains such as `vote.gov`, and real-time tracking of the election results, through domains such as `livevoterturnout.org`. This community also has overlaps with the 'March for Trump' community, where it was observed that users would share more links to domains such as `trumpmarch.com` and `joecheated.com` as it became more apparent that Trump had lost the election. Additionally, the co-occurrence network shows that archival sites, such as `archive.is`, were frequently posted to share links to archived reports of voter ballot errors, as well as blogs discrediting reports from so-called 'liberal' media channels.

As with the Reddit dataset, the URL co-occurrence graph for the 4chan dataset shows that the largest community of URL co-occurrence within election-related

posts from 4chan would share links to Trump-supporting domains showcasing his achievements, including `magapill.com` and `armyfortrump.com`. This community also frequently shared links to repositories on `pastebin.com`, where archives of Trump updates, interviews, campaigns and achievements were listed. These URLs are represented by the biggest nodes within the network graph in Figure 6.3.2, and were previously shown to be the most frequently posted domains in the 4chan dataset in Section 6.1. Again, another major theme in content-sharing behaviours on this platform was contesting the results of the election through posting links to statements from Trump and other political figures claiming the election was rigged. The co-occurrence network shows that there is significant overlap between these two communities, as demonstrated by the high number of connections between both sets of nodes.

The Stormfront URL co-occurrence network displays the least number of nodes in comparison to the other platforms. However, after applying the Louvain algorithm, it is apparent that there is a high level of modularity amongst the identified communities, as exhibited by the high edge-density within each cluster. Similar to the Reddit and 4chan network graphs, the URL co-occurrence network graph for election-related posts from Stormfront shows that contesting the election results and reporting on voter fraud were the most prominent themes in content-sharing on this platform. This can be observed from the two largest communities in the network graph. As well as this, the Stormfront posts would occasionally mass-post links to new-Nazi propaganda and antisemitic narratives. This mainly includes links to white-supremacy-related books on amazon and other digital libraries, in addition to archived blogs and manifestos.

6.3.3 Identifying Networks for Domain Sharing

To address RA3, any common domain-sharing activity linked across the four platforms was identified through further network analysis. In particular, domain-sharing network graphs were created by mapping links between the platforms and the most frequently shared domains to gain insight into some of the cross-platform

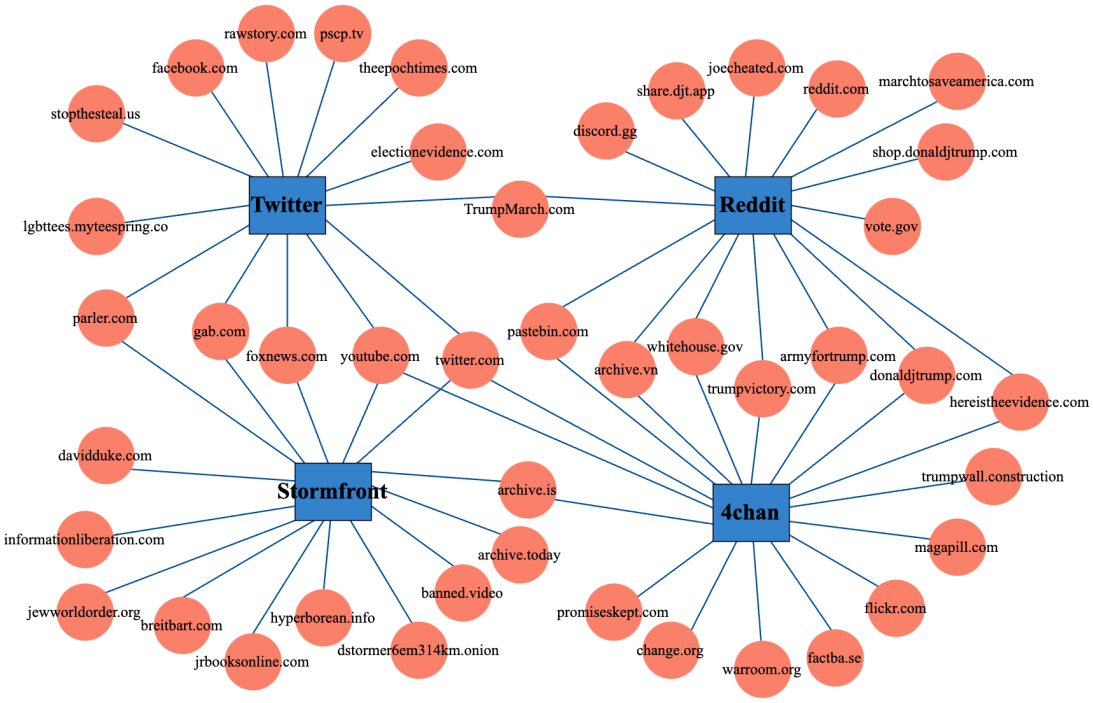

Figure 6.2: A domain network graph of the most-posted domains on each platform during the 2020 US election.

content diffusion dynamics. The resultant domain network is shown in Figure 6.2. From this, it can be observed that both 4chan and Reddit frequently promoted several of the same Trump-supporting domains, including `armyfortrump.com` and `trumpvictory.com`. These were posted much more frequently in the Reddit dataset and, as is evident in the URL analysis in Section 6.3.1, the majority of the most-commonly posted domains on Reddit were such Trump-supporting domains. Similar domains were also shared in the Twitter posts, particularly in the lead up to the Capitol riots, for instance `TrumpMarch.com`.

Both the Reddit and 4chan datasets additionally posted links to evidence of alleged "voter fraud", including `hereistheevidence.com`. In contrast, Stormfront was the most distinct out of all the platforms in that none of the most frequently posted domains found in the domain network were explicitly Trump-supporting. Instead, the majority of the domains shared were various blogs or websites hosting extremist literature linked to their white-supremacist identity. However, the shared-domain network graph displays how Stormfront users still

expressed support for Trump and participated in discussions related to Trump-initiated narratives like 'Stop the Steal' by sharing links to right-leaning news reports, including `breitbart.com`. The 4chan posts would additionally share multiple archival repositories, some of which were also posted on Reddit (`archive.vn` and `pastebin.com`) and Stormfront (`archive.is`).

The network graph in Figure 6.2 also showed that both the Twitter and Stormfront datasets would frequently share content from other social media platforms, including Parler, Gab and Facebook, especially to promote popular Trump-supporting accounts. Moreover, all four of the platforms would post links to content from Twitter and YouTube. Upon further inspection, it was found that Twitter often served as a news source on these platforms, mainly because most news publications and politicians have verified accounts here, where such accounts often post content on Twitter before any other platform. Similarly, Reddit users also frequently posted links to discord group chats, showing how hate groups may sidestep content moderation by redirecting their audiences to uncensored and encrypted platforms.

The insights into the content dissemination and diffusion patterns across all four platforms which can be derived from the shared-domain network graphs generally show that each platform has different diffusion dynamics, and promotes domains that are more specific to the platform and its user-base. Despite constituting only a small fraction of the total shared domains, this cross-platform analysis highlights similar domain-sharing activities that can still be observed. For instance, similar types of Trump-supporting domains were frequently shared between Reddit and 4chan users. There is also a lot of overlap between the types of domains that are shared across Twitter and Stormfront, where both platforms would often share URLs to content from other social-media platforms as well as far-right news sources.

6.4 Case Study 2: COVID-19 pandemic

6.4.1 URL Analysis

To address RA1, the cross-platform analysis is used to analyse the most frequently posted domains across all four platforms over the course of the COVID-19 pandemic,

where the 10 most popular domain names have been listed in Table 6.2. After extracting all of the URLs included within each dataset of posts, it is immediately apparent that the number of URLs being used within the Twitter dataset is significantly larger than those used within the other three datasets. In this case, the Twitter dataset included a total of 1,361,576 URLs, of which 243,771 were unique. From calculating the ratio of Twitter posts that made use of a URL to share content, it was found that around 99% of the Twitter posts included a URL.

This is similar to the findings from Case Study 1, where around 98% of election-related Twitter posts included a URL, suggesting that URLs are an integral aspect of content-sharing on Twitter. Previous studies have also determined that users often post their content on external websites to disseminate longer narratives due to the character limit on such platforms[290]. According to Kuzma et al. [291], these external websites are often shared along ideological or political lines on platforms such as Twitter, which is confirmed within this chapter.

In contrast, the 4chan, Reddit and Stormfront datasets made use of considerably fewer URLs. Reddit and 4chan users in particular made use of a similar ratio of URLs within their posts. The 4chan dataset included 182,388 URLs in total, with 49,862 of these being unique, and around 22% of the 4chan posts including a URL. The ratio of URLs within the Reddit posts was also around 23%, where a total of 10,709 URLs were found within the dataset, with 6814 of these being unique.

Interestingly, when comparing these findings to those of Case Study 1, it is clear that the usage of URLs in the Reddit posts is significantly higher in the election-related posts than the COVID-19-related posts. In addition to having a higher ratio of URLs used within the posts, a much smaller proportion of unique URLs were also shared over the course of the election. More specifically, within Case Study 1, Reddit users would frequently share the same links within their posts, though this does not occur as much in the COVID-19-related posts. It can therefore be inferred that the usage of URLs within Reddit posts is largely context-specific.

Similar to the findings from Case Study 1, this analysis found Stormfront posts related to the COVID-19 pandemic included the least amount of URLs. Here, a

total of 644 URLs were found in the Stormfront posts, with 156 of these being unique. As with the election-related dataset, around 5% of COVID-19-related Stormfront posts used a URL. These findings suggest that URLs do not play any significant role in the content-sharing behaviours of Stormfront users.

Table 6.2: The most-posted URL domains and their frequency across all four COVID-19-related datasets.

Twitter	**Reddit**	**4chan**	**Stormfront**
twitter.com: 15,007	archive.is: 273	pastebin.com: 4171	twitter.com: 51
zerohedge.com: 3215	youtube.com: 192	archive.is: 1532	youtube.com: 32
petitions.whitehouse.gov: 3010	thegatewaypundit.com: 142	promiseskept.com: 1655	redstatenation.com: 23
vocal.media: 2932	worldometers.info: 121	magapill.com: 1387	dailymail.co.uk: 22
washingtonpost.com: 2674	twitter.com: 115	twitter.com: 1086	thegatewaypundit.com: 18
youtube.com: 2099	statista.com: 107	cbp.gov: 868	militarytimes.com: 17
parler.com: 1906	reuters.com: 103	youtube.com: 732	timetobefreeamerica.com: 15
foxnews.com: 1764	washingtonpost.com: 96	zerohedge.com: 532	gab.com: 11
click2houston.com: 1046	vox.com: 87	factba.se: 492	worldometers.info: 9
cbsnews.com: 986	pastebin.com: 76	change.org: 457	benjerry.com: 6

Through conducting further analysis of the types of domains and subdomains that were posted the most within each of the four datasets, it can be seen that all four platforms mostly posted links to various news sources regarding updates with the COVID-19 pandemic. For instance, such domains, including `reuters.com` and `statista.com`, made up around 53% of the total URLs shared in the Reddit dataset. The Reddit posts would also frequently post Trump-supporting domains, such as `magapill.com`, and archival sites, such as `archive.is`, which made up 17% and 13% of the total URLs posted on Reddit during the pandemic.

Notably, the URLs extracted from the 4chan posts mostly consisted of Trump-supporting domains, such as `magapill.com` and `promiseskept.com`, which made up 35% of the URLs shared. Again, news reports regarding developments with the pandemic, including `zerohedge.com`, formed 23% of the URLs, whilst archival sites like `archive.is` made up 16% of the total URLs. Again, these archival sites were used to share content from preserved websites and reports consisting of disinformation and other controversial material to promote hateful and false narratives [206, 286].

Similar to the Reddit posts, the Twitter dataset mostly posted various news sources, the majority of which were right-leaning, such as `foxnews.com` and `zerohedge.com`, and often promoted misinformation regarding the pandemic. As with the findings from the URL analysis in Case Study 1, sharing links to content from other platforms made up around 15% of the total URLs, including domains for `parler.com`. In addition to this, the Twitter dataset was also shown to frequently share various petitions to the government, including to prevent trading with China and halt funding to the WHO and UN — such petitions formed 23% of the total URLs shared on Twitter. This observation is supported by previous findings from Phadke and Mitra, where hate groups were found to use Twitter to demand policy changes to negatively affect their targets [26].

Like with most of the other COVID-19-related datasets, the Stormfront posts mainly shared domains to news and updates related to the pandemic, where such domains formed 29% of the total URLs. The URLs shared within this dataset also consisted of links to content from other platforms, such as `gab`, as well as blogs promoting white-supremacist ideologies. These domains made up 24% and 35% of the total URLs, respectively.

6.4.2 Analysing the Presence of URL Co-Occurrence

The next stage of analysing the network dynamics of content-sharing activity amongst hateful users on all four platforms involved exploring URL co-occurrence behaviours. After applying the Circle Pack visualisation algorithm and the Louvain approach to community-detection, network graphs of any existing URL co-occurrence were created to identify common themes, or communities, of content sharing through URLs within each dataset, as shown below in Figure 6.3. The findings from this analysis were used to address RA2.

Overall, this analysis of the URL co-occurrence in the COVID-19-related datasets provided similar insights as those gained in Case Study 1. Again, Twitter was shown to have the most nodes, but the smallest thematic communities amongst these nodes. URL co-occurrence was mainly found in this dataset amongst posts

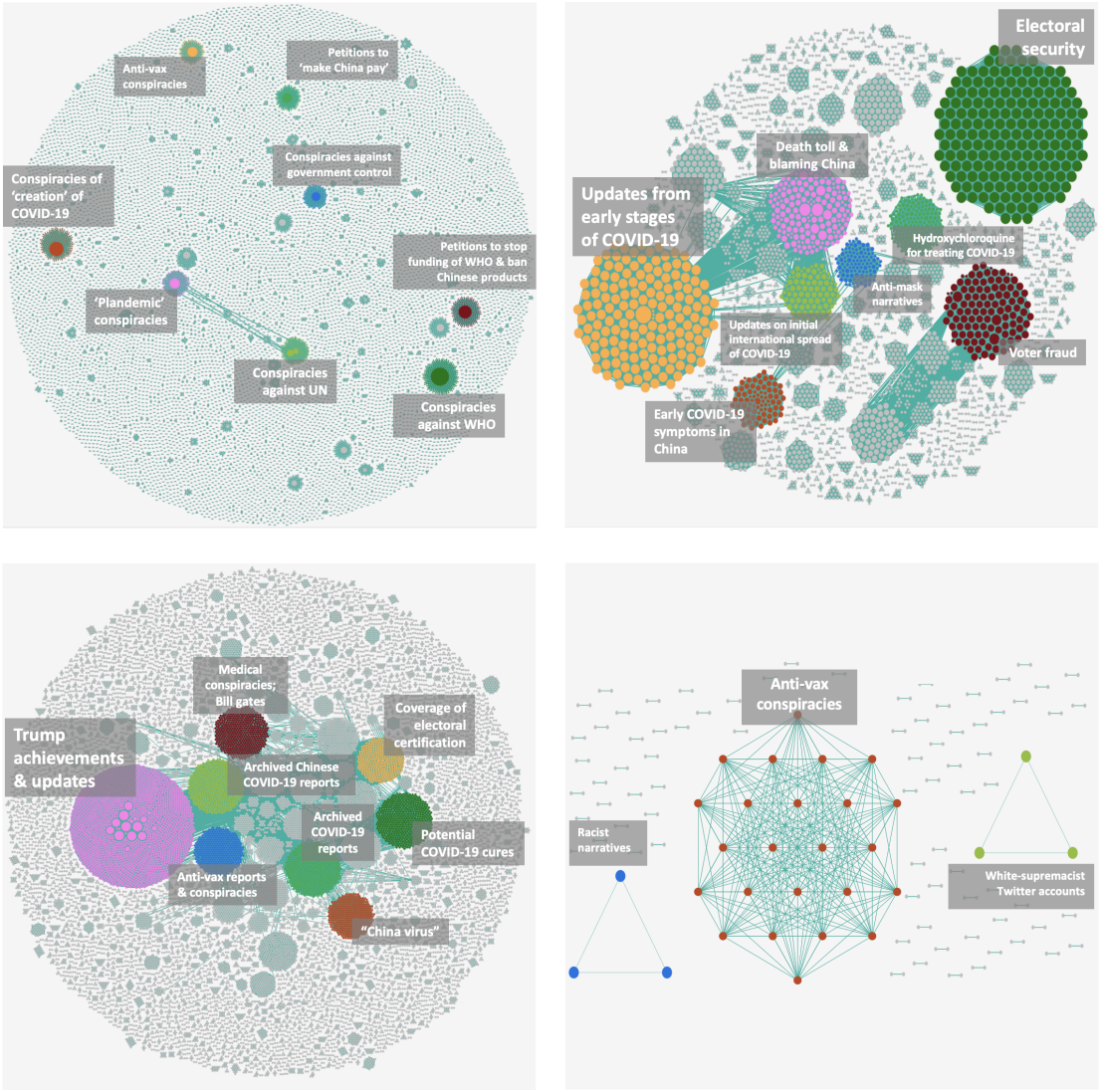

Figure 6.3: Network graphs of the URL co-occurrence behaviours found within each COVID-19-related dataset: Twitter (top-left), Reddit (top-right), 4chan (bottom-left), and Stormfront (bottom-right).

sharing various conspiracies about COVID-19. These include conspiracies involving the intentional 'creation' of the virus in labs in China, as well as conspiracies against the UN, the WHO and several governing bodies for imposing controlling measures. As mentioned in previous analyses in this chapter, this URL-posting behaviour could largely be attributed to the small character limit of Twitter posts, restricting users from sharing multiple URLs within the same post.

The network graphs for URL co-occurrence in Reddit and 4chan were similar in structure to those in Case Study 1 as well, where large communities of co-occurring

URLs were found within both datasets, with several communities exhibiting connections between them to depict overlapping themes. Many of these communities also exhibit a higher modularity, where the density of connections within these communities is higher than the other platforms. The 2020 US election was a major theme in content-sharing activities on both Reddit and 4chan, showing that COVID-19-related posts would frequently be discussed within the context of the election. This is not surprising given that the political response to the pandemic was an integral aspect of political manifestos and campaigns during the election. In addition to sharing links to various news sources regarding updated case numbers and death tolls from COVID-19 (such as `worldometer.com` and `statista.com`), both Reddit and 4chan posts also shared reports and blogs promoting anti-mask and anti-vax conspiracies, as well as conspiracies blaming China for "manufacturing" the virus.

One major difference in the URL-sharing behaviours from both these platforms is that the majority of URL co-occurrence in Reddit posts can be found within these communities of high modality, which were identified through the circle pack and Louvain algorithms. In contrast, 4chan users would frequently post URLs outside of these communities as well. This implies that Reddit users would largely share URLs that aligned with particular topics of discussion, whereas 4chan users would share content unrelated to major topics of discussion as well. This could be attributed to the 'online community' structure of the r/donaldtrump subreddit, where the Reddit posts were collected from.

The COVID-19-related posts collected from Stormfront show the least amount of URLs in the co-occurrence network graph, which was also found to be the case in previous analyses within this chapter. URL co-occurrence mostly takes place within the largest community identified by the Louvain algorithm, which is anchored in promoting anti-vax conspiracies. Here, Stormfront posts were found to share links to various blogs, news reports and pseudo-studies, including those from the domains `timetobefreeamerica.com` and `theglobeandmail.com`, which promoted the idea that vaccines were being used to control and change humans. In comparison to Case Study 1, the Stormfront URL co-occurrence network is

significantly smaller with COVID-19-related posts than for the election-related posts discussed in Section 6.3.2. This suggests that content-sharing practices through the usage of URLs is largely context-specific on Stormfront.

6.4.3 Identifying Networks for Domain Sharing

The final aspect of applying the cross-platform analysis framework to the study of hateful networks harnesses SNA techniques to understand how COVID-19-related content diffused across all four platforms through domain-sharing network analysis. Again, this involves creating domain-sharing network graphs by mapping links between platforms and frequently shared domains in order to gain insight into the dynamics of cross-platform content diffusion. The domain network graph for COVID-19-related content is shown in Figure 6.4. This analysis revealed that news sources were the most commonly shared domains across all of the platforms. These news sources were primarily used to promote various conspiracy theories about the pandemic, the source of the virus, and anti-vax and anti-mask narratives. The mix of the types of news sources shared by users, however, varied across platforms.

Within this case study, Twitter and Reddit shared a mix of far-right and less controversial news channels, such as `washingtonpost.com`, whereas more underground platforms like 4chan and Stormfront primarily shared far-right news sources and blogs, such as `thegatewaypundit.com`. False conspiracies blaming China for COVID-19 and expressing frustration with China and other prominent political figures were also prevalent across all platforms. The shared-domain network graph in Figure 6.4 shows that while such content is more prevalent on 4chan and Stormfront, it is still disseminated across multiple platforms, whereas other less controversial news sources diffuse much more frequently through users within the same platform, like Twitter and Reddit.

As with Case Study 1, content from Twitter and YouTube are frequently shared by users across all four platforms, where such content was used as news sources themselves. This includes statements from various political figures and video accounts of various events, such as reports from COVID-19 patients from hospitals

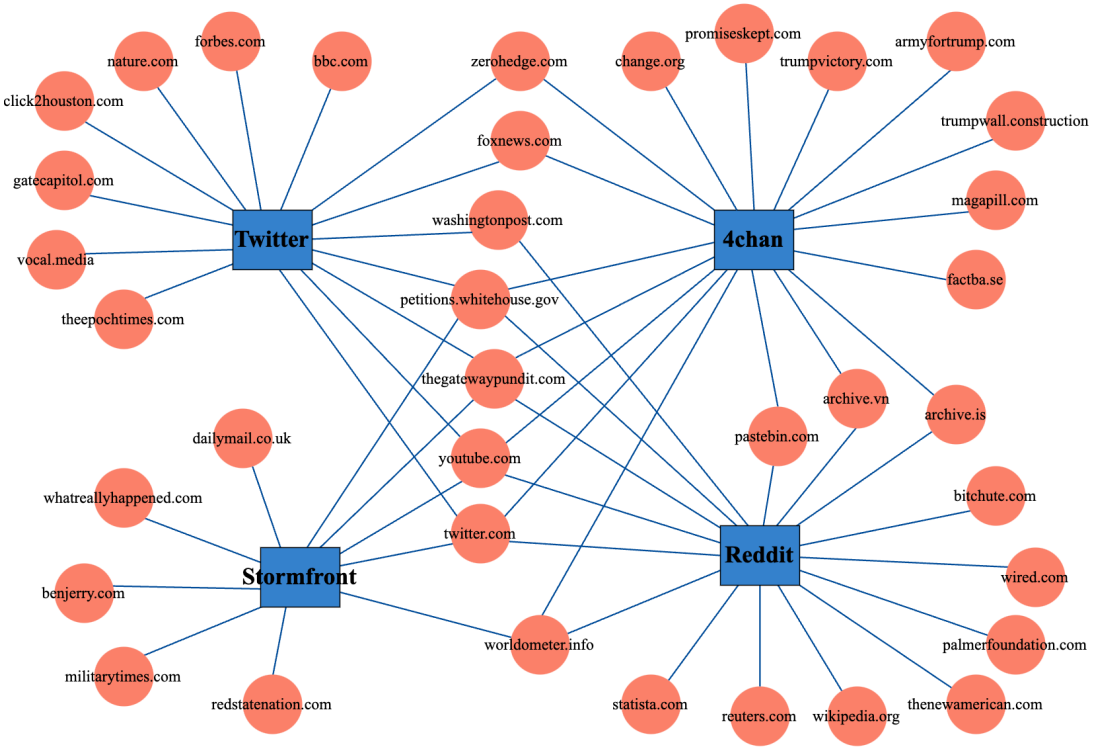

Figure 6.4: A domain network graph of the most posted domains on each platform during the COVID-19 pandemic.

in China. Overall, content from these two platforms diffused much quicker and more frequently through cross-platform networks of hateful users. Users on all platforms also frequently circulated petitions to government officials to place political sanctions on China and halt funding to organisations such as the WHO, who were falsely believed by a majority of the users to have intentionally manufactured the virus. This suggests that such content diffuses much more frequently across all platforms when users are working towards a common goal.

In addition to this, the domain-network graph shows that 4chan, Reddit and Stormfront posts shared similar content from domains including `worldometer.info`, which provided live tracking of updates for case numbers and death tolls related to COVID-19 across various jurisdictions. This is consistent with previous findings, where discussions centred around developments in the pandemic and death tolls were shown to be a major topic of discussion on Reddit and 4chan in Chapter 5. Users from both 4chan and Reddit would, again, share similar archived reports

spreading false narratives around the pandemic, and data repositories from domains including `archive.is` and `pastebin.com`. This indicates that the 4chan and Reddit datasets generally exhibited similar content diffusion dynamics, where it is clear that more of the same content is shared across users from both these platforms. This is also consistent with findings from previous research conducted by Vosoughi, Roy and Aral [32], where the authors found that false rumours spread faster and more broadly than their true counterparts.

Similar to the findings from Case Study 1, the 4chan dataset was again shown to predominantly promote Trump-supporting domains, including `magapill.com` and `promiseskept.com`, where this content-sharing behaviour was mostly unique to 4chan within the context of COVID-19-related content. Since this was also observed in Case Study 1, this indicates that Trump support is a much more integral aspect of content shared on 4chan beyond election-related content in comparison to the other platforms.

Another observation that can be made from this shared-domain network analysis is that links to other social-media platforms were found to be less prevalent in the Twitter dataset for this case study than the election-related Twitter posts. This suggests that content is not always disseminated through other platforms on Twitter unless there is a shared audience users are trying to address, or a shared movement that users are trying to promote.

6.5 Summary

In this chapter, the cross-platform analysis framework was harnessed to explore how content-sharing behaviours through the posting of URLs compare across different online platforms, and how this can provide insights into cross-platform network and content diffusion dynamics. In particular, this involved applying various network-analysis techniques to data collected during the 2020 US election and COVID-19 pandemic from Twitter, Reddit, 4chan and Stormfront. URLs were first extracted from all the datasets to compare the types of content most frequently promoted on each platform, where it was apparent that URLs were used regularly across

most platforms to promote social movements and narratives, such as the "Stop the Steal" movement during the US election. It was also clear that URLs played an integral role in promoting content on platforms like Twitter, where the character length of posts is limited, but significantly less on more underground platforms like Stormfront. This demonstrates that platform-specific constraints and features impact the way users share content.

Through analysing the URL co-occurrence on each platform in both case studies, this chapter details how, despite URLs being used the most in Twitter posts, very little co-occurrence existed amongst the shared URLs in the Twitter datasets. Reddit and 4chan, on the other hand, exhibited considerably high amounts of URL co-occurrence, which usually occur in large communities of shared themes. This indicates that URLs are commonly posted on these platforms as a way to boost certain campaigns and narratives, such as anti-vax conspiracies related to the COVID-19 pandemic. This implies that users on these platforms might post URLs to promote and reinforce particular viewpoints or conspiracy theories.

Finally, networks graphs of shared domains across all four platforms were created to gain insight into cross-platform content dissemination patterns. An observation made from this analysis is that users from Reddit and 4chan would frequently post similar types of content from the same domains, especially archived reports and repositories, thus implying a higher diffusion rate between both these platforms in particular. These findings from applying the cross-platform framework, therefore, shed light on the dynamics of information sharing, content dissemination, and thematic communities within different online platforms.

The next chapter validates the findings gained from applying the cross-platform analysis framework for online hate to various case studies.

7

Validation of the Cross-Platform Analysis Framework

7.1 Introduction

The previous chapters in this thesis have posited how a cross-platform analysis framework can be used to gain novel and unique insights into the online behaviours of hateful groups and users. However, it is important to reflect on the validity of any analysis framework in order to assess its efficiency at capturing the phenomenon they are designed to analyse, as well as the value of the insights it provides. One way in which this can be done is through the use of case studies to evaluate the practical applications of the analysis framework to real-world scenarios [215].

This validation process can provide several insights into the effectiveness of the cross-platform analysis framework. For instance, it can highlight the strengths and weaknesses of the framework in identifying patterns and behaviours across different platforms. By analysing real-world examples of online hate incidents, the framework can be evaluated according to how well it captures the dynamics of online hate, and identify areas where the framework may need to be refined. Validation of the framework can also identify the limitations of the findings that can be obtained from applying the cross-platform analysis framework. Through comparing the results generated by the framework to ground truth data, it is possible to identify areas

where the framework may be less effective, such as in generalising hateful behaviours exhibited on online platforms across various contexts.

Within the context of validating the cross-platform analysis framework for online hate research developed in this thesis, Chapter 5 and Chapter 6 have detailed how two primary case studies, the 2020 US election and the COVID-19 pandemic, were used to demonstrate the applicability of the framework. The validation process of this cross-platform analysis framework thus involves assessing the performance of the framework against a list of validation criteria, which includes reflecting on the insights gained from applying the framework to various case studies. So as to demonstrate that this framework can be applied beyond the two principal case studies used in the previous chapters of this thesis, an additional test case study is also used here to illustrate how online-hate researchers would be able to make use of the framework in practice.

The list of the validation criteria that have been used to evaluate the validity of the cross-platform analysis framework developed in this thesis has been adapted from the framework requirements detailed in Chapter 4, as well as the validation measures outlined in [292]. These criteria are listed below:

1. **Clarify the theoretical foundations of the framework:** Clarify the theoretical foundations of the framework, including the key concepts and assumptions that underlie it. This is done by conducting a thorough review of the existing literature on online hate and related concepts.

2. **Identify the relevant contexts:** Identify the social, cultural, and political contexts in which the framework is intended to be applied. This helps to ensure that the framework is relevant and applicable to the specific context in which it will be used.

3. **Evaluate the scope of the framework:** Evaluate the scope in terms of the ability of the framework to capture different forms of online hate. This can be done by reviewing the definitions of online hate used within the framework

and examining the applicability of the framework to different types of online content.

4. **Assess the practical applications of the framework:** Assess its practical applications, including the data collection, processing, and analysis techniques it uses. This is done by applying the framework to relevant datasets and case studies and evaluating its ability to provide additional insights into the cross-platform behaviours of hateful users.

5. **Consider ethical concerns:** Take into account the ethical considerations related to the use of the framework, including issues of privacy, confidentiality, and data protection. This is done by conducting an ethical review of the framework and ensuring that it adheres to relevant legal and ethical guidelines.

The remainder of this chapter is structured as follows. Section 7.2 discusses the validation of the cross-platform analysis framework using the above validation criteria through the context of the findings gained from Case Study 1 (online hate during the 2020 US election) and Case Study 2 (online hate during the COVID-19 pandemic). This will include reflecting on the novel insights gained from applying the framework to these particular case studies, particularly when compared to the extent of the findings from single-platform approaches.

Section 7.3 will then detail an additional validation exercise, where the cross-platform analysis framework is applied to another test case-study, so as to demonstrate how it can be used by online-hate researchers. The findings from this study will be used to further assess the practical applications of the framework, which is a key component of the validation criteria. Finally, this chapter reflects on the insights gained from the validation of the cross-platform analysis framework, as well as any limitations and implications in Section 7.4.

7.2 Validity of the Framework

7.2.1 Clarify the Theoretical Foundations of the Framework

The analysis framework proposed in this thesis has been designed to provide novel insight into the cross-platform behaviours of hateful groups and users. This is based on the assumption that hateful groups and users make use of multiple platforms to promote hateful narratives. Through conducting an extensive literature review of academic research within this field, several studies were found to provide evidence showcasing online hate across multiple platforms. Previous research has also studied a variety of different platforms individually, including both mainstream and underground platforms, further proving the presence of hateful discourse over multiple platforms, though such platforms have rarely been explored together.

This thesis first establishes a definition of online hate by combining key elements from previous literature [23], governing bodies [37], NGOs [38], and Platform terms and conditions [47]. More specifically, this definition was determined with regards to four particular dimensions: online hate has specific targets; online hate is to incite violence or hate; online hate is to attack or diminish; and humour has a specific status. This definition was used, along with accounts from previous works, to outline which behaviours, users and environments constituted as hateful.

This thesis also presents a conceptual model for online hate in Chapter 4, which includes the major components pertinent to online-hate research; namely the causes and consequences of online hate, and the methods that can be used to examine these. Chapter 4 discusses how the cross-platform analysis framework contributes to research in the field of online hate by providing a novel approach to the analysis of online hate, so as to help address its causes and consequences. Through comparing existing frameworks for online hate analysis within the literature review, gaps within the research landscape were identified and used to define the requirements of the cross-platform analysis framework. The insights from this aided the design of the functional components of the framework.

7.2.2 Identifying the Relevant Contexts

Chapter 4 defines how the framework is intended to be applied to the study of hateful content across multiple online platforms. This framework is designed to study groups from different hateful ideologies, including far-right and white supremacist ideologies. Depending on the type of hate wanting to be analysed, specific contexts for analysis can be chosen by exploring reports of hateful incidents or providing evidence for an increase in hateful content. For instance, in this thesis, the framework was used to analyse the hateful discourse and behaviours of users and groups from white-supremacist platforms or accounts. The framework was thus applied to data collected from the 2020 US election, where this type of hate was shown to be prevalent by various news reports and other academic literature [208]. Similarly, the COVID-19 pandemic was another context within which the prevalence of hateful narratives increased [202].

7.2.3 Evaluate the Scope of the Framework

Since the framework is designed to identify hateful content from multiple platforms, it is not restricted to the study of any particular type of hate, as mentioned previously, or any particular platform. However since this framework has mainly been developed based on insights from English-language-centric research and contexts, the framework is used with English-only content within this thesis. Though this does not explicitly restrict the application of this framework to content from other languages, further refining and adjustment would perhaps be required to adapt the design and functionalities of the framework to such content. Additionally, this framework has been developed with analysis techniques that are predominantly applicable to textual content. Since previous literature has established on many occasions that online hate can be spread through various forms of content, such as graphic content (particularly through the usage of memes [293]), this does limit the insights that could be gained into the comparison of hateful content across multiple online platforms.

7.2.4 Assess the Practical Applications of the Framework

The practical applications of the cross-platform analysis framework have been assessed through its use with various case studies, and its ability to provide additional, novel insights to the cross-platform behaviours of hateful users that would not be provided by a single-platform approach. This thesis particularly carries out analysis using the framework in two case studies: the 2020 US election and the COVID-19 pandemic. For both these case studies, data was collected from the four platforms Twitter, Reddit, 4chan and Stormfront by using platform APIs, data-capturing toolkits, and publicly-available datasets collected by other researchers. The datasets for both case studies were analysed using various analysis techniques, including NLP and machine-learning techniques to extract the most discussed topics on each platform, as well as network-analysis approaches, such as analysing URL co-occurrence networks and shared-domain networks, to gain insights into the content-sharing dynamics of hateful users across different platforms. Some of the key findings from this analysis of both case studies are provided below.

Case Study 1: 2020 US Election

Through using various NLP methods for analysis, the cross-platform analysis framework was able to highlight similarities and differences between the hateful discourse during the 2020 US elections in each of the four platforms. More specifically, this analysis found that the majority of the discussions on each platform centred around promoting social movements and narratives protesting the results of the election, such as 'Stop the Steal' and 'March for Trump'. Twitter posts in particular were found to share details for rallies being organised by protesters and encouraging other users to join. The Reddit posts were similarly observed to campaign heavily for other users to partake in certain actions, including voting for Trump in the lead up to the elections. Generally, both these more mainstream platforms were found to influence and encourage their audience to join certain social movements, where further linguistic analysis confirmed these observations

in the presence of more persuasive techniques being used within these posts, such as the usage of certain pronouns.

The cross-platform analysis also found that the more underground platforms like 4chan and Stormfront heavily promoted narratives related to the "Us vs. Them" dichotomy mentality, where derogatory language is often used to describe groups of "others", especially for black people and Jews. This was indicated by the higher number of plural pronouns which were used to foster a stronger sense of "white identity" amongst users on these platforms. Similarly, analysing the general sentiment of the posts from each dataset showed that 4chan and Stormfront posts displayed a more negative emotional tone, whereas Twitter and Reddit displayed a more positive tone.

Through harnessing network-analysis techniques, the cross-platform analysis framework also provided novel insight into the content-sharing practices on all four platforms. For instance, the analysis shows that URLs are used much more frequently in the Twitter and Reddit election-related datasets to share content. Reddit posts in particular would use these to promote certain narratives, like voter fraud, or encourage other users to take specific actions, like register to vote. Stormfront posts were found to use URLs the least, indicating that URLs did not play as important of a role in sharing content on this platform.

Using the analysis framework to study the URL co-occurrence on each of the four platforms revealed how platform affordances can impact content-sharing behaviours. Although URLs were shown to be a major part of content sharing on Twitter, the small character limit allowed for posts on this platform also meant that multiple URLs could not always be shared within posts, as indicated by the very little co-occurrence observed in this dataset. Platforms like Reddit and Stormfront, on the other hand, exhibited a lot of URL co-occurrence. Applying community-detection algorithms to these network graphs showed that the majority of the URLs shared within these datasets fit into particular themes and narratives, suggesting URLs were primarily used on these platforms to support and propagate hateful discourse.

Finally, additional network-analysis techniques were utilised to create and analyse network graphs of shared domains across all four platforms. Through this, it was found that Reddit and 4chan users in particular would often promote content from the same domains, as a lot of content is shown to be shared across both these platforms. This includes various archived reports and repositories, suggesting there is a lot of overlap between the user-base or audience on these two platforms.

Case Study 2: COVID-19 Pandemic

Using the cross-platform analysis framework to apply various NLP techniques to data collected from Twitter, Reddit, 4chan and Stormfront during the COVID-19 pandemic also provided unique insight into the type of content promoted on different platforms. For instance, this analysis confirmed some of the observations made in Case Study 1 regarding the posting behaviours across all four platforms. Specifically, offline developments were shown to have more of an impact on the frequency of posting from users on platforms with larger audiences, like Twitter, Reddit and 4chan, where offline developments would trigger peaks in posting. In contrast, Stormfront exhibited more steady and consistent posting behaviour in both case studies, indicating that such events do not impact discourse on this platform as much.

Through using the cross-platform analysis framework to compare major topics of discussion on each of the platforms, it was observed that false narratives and various conspiracies surrounding the pandemic controlled much of the discourse on all four platforms. In particular, these included blaming China and different governing bodies for the intentional manufacture of the virus to control populations. As with Case Study 1, the more underground platforms like 4chan and Stormfront were shown to frequently promote the "Us vs. Them" dichotomy mentality, where, within the context of COVID-19-related discourse, this was used to promote false and hateful narratives related to various groups of "others", including immigrants and Jews, causing and intentionally spreading the virus. This is again exhibited through the higher presence of third-person plural pronouns.

Harnessing network-analysis techniques to examine the content-sharing behaviours of users on all four platforms through the usage of URLs confirmed many of the observations made in Case Study 1. Namely, this included how URLs were an integral aspect of content sharing on Twitter, largely due to its small character limit, and, again, very little co-occurrence was found in Twitter posts. Interestingly, Reddit users were shown to use URLs significantly less in COVID-19-related discussions as compared to Case Study 1, which could be attributed to the fact that there was much less campaigning for certain movements, like voting for Trump and 'Stop the Steal'. This suggests that certain content-sharing behaviours, like the usage of URLs in posts, are context-specific for platforms like Reddit.

Through examining the network graphs created to display URL co-occurrence on each of the four datasets, Reddit posts were again shown to exhibit the most URL co-occurrence. The findings from this analysis also confirmed observations made in Case Study 1, where URLs shared within Reddit posts fit into larger overall themes, suggesting that URLs are primarily used on this platform to support and promote particular narratives.

Applying the cross-platform analysis framework to the study of the diffusion and dissemination dynamics of content across multiple online platforms also provided novel insight to online hate research. More specifically, network-analysis techniques were used to show how the most commonly shared content through URL domains across all four platforms was right-leaning news sources. Like previous findings detailed in this thesis, 4chan and Reddit posts displayed the most content sharing between users, where archived articles and repositories were frequently shared on both platforms. As mentioned previously, this could indicate overlaps in the user-base or audience of both these platforms.

7.2.5 Consider Ethical Concerns

Various ethical considerations were taken into account during the development and application of the cross-platform analysis framework. Specifically, Chapter 3 details how the Central University Research Ethics Committee at the University of

Oxford was consolidated to obtain relevant ethical approvals prior to any research being carried out. Despite all of the data collected and used over the course of this research being publicly available, the necessary precautions were taken to anonymise the data as much as possible, so as to prevent the publication of any identifiable data. Furthermore, university guidance was followed throughout the research process regarding the usage of security-sensitive research material. Counselling services provided by the university were also consulted to ensure the psychological safety of the main researcher.

7.3 A Test Case Study for Validation

In order to further assess the practical applications of the cross-platforms analysis framework, it is applied to an additional test case study to demonstrate how online-hate researchers would be able to make use of the framework in practice. This will also confirm that the framework is applicable beyond the two primary case studies detailed in Chapter 5 and Chapter 6. As with the previous case studies, the test case study is used to validate the framework with particular emphasis on the additional and novel insights it is able to provide in comparison to traditional, single-platform approaches.

The test case study used within this validation of the cross-platform analysis framework is described as follows:

> *Comparing the hateful behaviours and narratives during the 2020 Black Lives Matter protests on Twitter and Reddit.*

Following the killing of George Floyd at the hands of a Minneapolis police officer on May 25th 2020, widespread protests emerged throughout the world, primarily led by the Black Lives Matter (BLM) movement. As well as an increase in support for the BLM movement during this period, disinformation and hateful narratives around the protests and protesters also spread rapidly on various online platforms, including Twitter and Reddit [294]. Previous reports have also discussed how this online hate and disinformation was intentionally used by far-right hate groups

to spread fear and division, whilst also undermining the efforts of protesters to bring structural change [295].

To analyse and compare the activity and discourse of hateful users during the 2020 BLM protests from Twitter and Reddit, the cross-platform analysis framework was used with particular regard to the following two research aims:

- **RA1:** Explore how the participation and posting trends compare on both platforms over the course of the protests.
- **RA2:** Identify and compare the main topics of discussion for hateful users on both platforms.

Further details on the data collection and the analysis methods used with the cross-platform framework are provided below.

7.3.1 Data Collection and Analysis Methods

To address the research aims, the analysis approach used in this test case study thus comprises two stages: (1) observing the posting frequency on Twitter and Reddit during the protests, and (2) conducting topic modelling on both datasets. As with Case Study 1 and Case Study 2 discussed in previous chapters of this thesis, Twitter data was collected using the Twitter API from the same 478 far-right and white-supremacist accounts previously identified through information provided by the Southern Poverty Law Centre (SPLC)[1]. The Reddit posts were, again, collected using the Pushshift API [13] from the r/donaldtrump and r/The_Donald subreddits, which were linked to spreading online hate over the course of the BLM protests [252, 294].

In order to gain a complete overview of the online discourse over the course of the BLM protests on both platforms, data was collected from 1st May to 1st August 2020. This time frame encompasses all the key events surrounding the protests, including the May 25th killing of George Floyd that initially sparked the protests, up until the end of July where protests had subsided in most parts of the US. A list of key terms related to the BLM protests was used during the

[1]https://www.splcenter.org/hate-map

collection process to ensure only relevant content was collected. These terms include "blm", "white lives matter", "all lives matter", "blue lives matter", "back the blue", "police lives matter", and "George Floyd".

The number of posts collected from both Twitter and Reddit are as follows:

- Twitter BLM dataset: 1,743,088 posts.
- Reddit BLM dataset: 136,071 posts.

After collecting both datasets, the participation trends were measured by observing the frequency of posts being shared weekly online over the course of the protests, in answer to RA1. The second stage of the analysis involved harnessing topic-modelling techniques to identify the main topics of discussion on each platform, so as to address RA2. The Latent Dirichlet Allocation (LDA) topic detection model was again applied to both datasets to provide a distribution of a selected number of topics. In this test case study, the number of topics specified for the LDA topic model was five, as this was the number of topics that were found to produce the most distinct topics in both datasets. The findings from this analysis thus identify the five most discussed topics on Twitter and Reddit during the 2020 BLM protests.

Before carrying out this analysis, the pre-processing steps outlined in Chapter 5 were again applied to both the Twitter and Reddit datasets, including removing duplicate posts, punctuation marks and platform-specific noise. To yield more accurate topics, all of the posts were tokenized and a TF-IDF array was created to fit the LDA topic model. As with the previous case studies, this analysis is carried out using the Pandas[2] data-analysis library provided by the Python programming language.

7.3.2 Results and Findings

Participation Trends

In answer to RA1, the analysis first explores the frequency of content being posted within each dataset over the course of the collection period during the BLM

[2]https://pandas.pydata.org/

protests. In Figure 7.1, it can immediately be seen that posting behaviours peaked significantly on both Twitter and Reddit in the aftermath of the May 25th killing of George Floyd. This peak in posting appears sooner in the Reddit dataset, where the frequency of posts increased in the immediate aftermath of the killing (week commencing May 25th). The Twitter dataset on the other hand exhibited a steep peak a couple of weeks later (week commencing June 8th). By this time, protests were being held in all states in the US, and increased clashing between police and protesters was being reported [296].

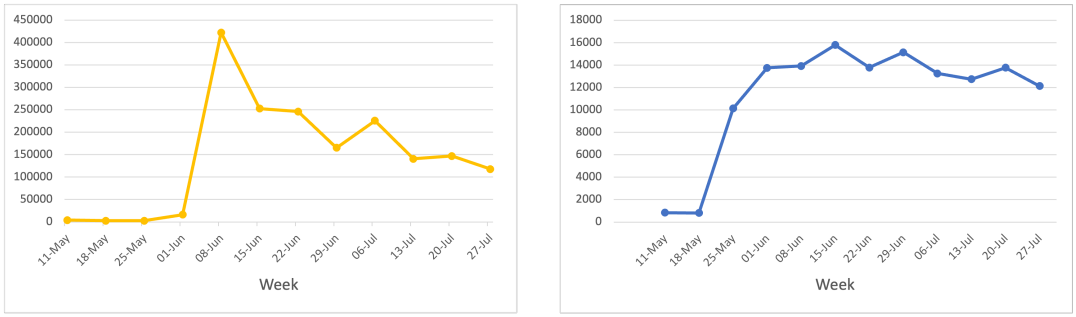

Figure 7.1: Graphs showing the frequency of posts across both datasets over the course of the 2020 BLM protests: Twitter (left) and Reddit (right)

Another key difference between the posting frequency exhibited on both platforms is that, whilst the frequency of Reddit posts remained mostly steady throughout the rest of the collection period, the post frequency on Twitter dropped significantly after the initial peak. This suggests that the BLM protests remained a prominent theme in the discussions on Reddit, where a significantly high volume of posts related to the protests can be seen over the following couple of months in the aftermath of the killing. In contrast, Twitter posts seemingly peaked considerably during the height of restrictions being put in place by law enforcement to control the protests, but then almost immediately decreased in frequency in the following months as protests began to subside.

Therefore, in answer to RA1, the participation trends across both platforms seem to be similar, in that they significantly peak during the height of the protests, showing that offline events majorly influence online discourse on both Twitter and Reddit, though these discussions are maintained on Reddit for longer.

Topic Modelling

A topic model, using the LDA topic detection approach, also provided further insight into the most discussed subjects within each dataset, with the five identified topics and the percentage of posts containing them being listed in Table 7.1. Again, this percentage was obtained by extracting the most dominant topic out of the five topics in each post, and a cumulative total of the number of posts for each topic was calculated and represented as a percentage of the total posts in the dataset.

Table 7.1: A topic model of the most discussed topics during the 2020 BLM protests and the percentage of posts containing them.

	Twitter	**Reddit**
Topic#1	battle anonymous, victory online, major victory, troops complete, complete decimation **(13%)**	blue lives matter, police lives, blm, black lives **(22%)**
Topic#2	blue lives matter, tribute, tribute fallen, cop attacked, fallen brother, blue lives matter tribute **(22%)**	antifa, marxist, terrorist, support blm, funding, f*** blm **(25%)**
Topic#3	whites pure, white lives matter, killing whites, pure racism, blacks killing, rioters **(19%)**	statistics, police shot, cop, cops killed, number, death, police **(17%)**
Topic#4	unarmed cops, killed cops, protests begin, killing, cops dead **(21%)**	defund police, want to defund, biden, biden defund, abolish, security **(13%)**
Topic#5	kpop, kpop stans, white lives matter, fancams, stans **(25%)**	black people, killed police, unarmed, police, blacks, riots, protesters **(23%)**

When comparing the topics from both the datasets, it can be seen that similar topics are generally discussed or mentioned. For instance, common topics on both Twitter and Reddit include the "blue lives matter" narrative in support of police officers in their clashes with BLM protesters, as well as reports on police officers attacked or killed during these clashes. Such topics dominate the majority of the discourse in both datasets, with 62% of all the posts in Reddit discussing these narratives (Topic#1, Topic#3 and Topic#5), and 43% of Twitter posts discussing these topics (Topic#2 and Topic#4).

The topic model also demonstrated how Reddit users would often spread false conspiracies around the BLM movement, namely how it was being funded by Antifa (a left-wing anti-fascist, anti-racist political movement), who they described as "marxist terrorists" (Topic#2). Such narratives were discussed in around 25% of the Reddit posts. Another topic identified by the topic model in the Reddit dataset included users complaining about the "defund the police" movement promoted by BLM protesters, and how this movement was being supported by Biden (Topic#4). Around 19% of Twitter posts were also found to attack the BLM movement by associating it to a racist attack on white or "pure" people, with frequent mention of white people being killed by black people and BLM protesters (Topic#3).

Interestingly, a huge portion (29%) of Twitter posts were also found to discuss K-pop-related topics (Topic#5), a topic seemingly completely unrelated to the BLM protests. Upon further inspection, it was noted that several accounts and hashtags associated with hate groups, such as '#whitelivesmatter', were hijacked by K-pop fans during the height of the backlash against the BLM protests, so as to take over hateful narratives by promoting K-pop-related content instead. As mentioned previously, such behaviour has been a frequent occurrence in the past few years, as online users, mainly on mainstream platforms like Twitter, have realised that they can break up hateful discourse online by taking over hashtags predominantly used to promote hate [297].

7.3.3 Summary

This test case study harnessed the cross-platform analysis framework to investigate how hateful content during the 2020 BLM protest compared across Twitter and Reddit, with a particular focus on the frequency of content being posted and the most discussed topics. The analysis reveals that both platforms exhibited a significant peak in posting behaviour in the aftermath of George Floyd's killing, with Reddit maintaining a high volume of posts related to the protests for a longer period compared to Twitter. The identified topics on both platforms include the "blue lives matter" narrative, reports on police officers attacked or killed during clashes,

false conspiracies spread by Reddit users, and attacks on the BLM movement. Interestingly, a major portion of Twitter posts discussed K-pop-related topics, which were shown to be hijacked by K-pop fans to take over hateful narratives.

Overall, the analysis suggests that offline events have a major influence on online discourse on both Twitter and Reddit, with discussions on Reddit being maintained for a longer period.

7.4 Discussion

In order to examine the strengths and weaknesses of any analysis framework, it is essential that a thorough validation is carried out. This chapter details the validation process used for the cross-platform analysis framework developed in this thesis. This involved assessing the framework against a list of validation criteria, which were adapted from previous studies as well as the framework requirements outlined in Chapter 4. In particular, the cross-platform analysis framework was evaluated by clarifying the theoretical foundations of the framework, identifying the relevant contexts in which the framework can be used, specifying the scope of the framework, assessing its practical applications, and finally discussing the ethical considerations related to the use of the framework.

This especially focussed on assessing the practical applications of the cross-platform analysis framework through the use of case studies. Previous chapters in this thesis have detailed its application to data collected from two primary cases — the 2020 US election and the COVID-19 pandemic. Through validating the findings gained from this analysis, this chapter highlights how using a cross-platform approach to analysing online hate can provide additional and novel insights that existing, single-platform approaches would miss. This includes the observation that contemporary social events occurring offline have much more of an impact on the online discourse on mainstream platforms like Twitter and Reddit, in comparison to smaller underground platforms like Stormfront. This is exhibited through both the posting frequencies on each platform as well as the major topics of discussion identified within the posts.

Another notable observation provided by the cross-platform analysis framework is the differences in content-sharing practices on each of the platforms, specifically through sharing URLs. This was found to be a much more integral aspect of content sharing on Twitter, where URLs were mostly used to support or promote particular narratives. Reddit posts, on the other hand, showed differences in such behaviour within both case studies, indicating that certain content-sharing practices on online platforms are context specific. Similarly, using network-analysis techniques to create and analyse networks of shared domains on each of the platforms highlighted how the Reddit and 4chan datasets promoted a lot of the same types of content, showcasing a higher content diffusion rate between these two platforms. This suggests an overlap in the audience or user-base of these platforms.

To further assess the applicability of the framework, this chapter details its application to a further test case study, which was used to demonstrate how online-hate researchers would be able to use the framework in practice. The test case study comprised of a comparative analysis of online hate during the 2020 BLM protests on Twitter and Reddit, with particular regard to the trends in posting frequency on both platforms, as well as the main topics of discussion in both sets of posts.This analysis also highlighted shared narratives that are promoted on both platforms, and differences in posting trends over the course of certain offline events.

Overall, the validation of the cross-platform analysis framework demonstrated how it is able to provide unique and novel insight into the study of online hate, which would be otherwise overlooked by existing single-platform approaches. The framework therefore provides a more comprehensive understanding of online hate. In addition to the strengths of the framework, the validation process also highlighted certain limitations within the design and application of the framework, mostly due to its scope. These are discussed in further detail in the next chapter.

8 Conclusions and Future Work

This thesis aimed to identify and address research limitations in the field of online hate by proposing a novel approach to analysing hateful content and behaviours on multiple online platforms. In this final chapter, the research findings are concluded to assess how they have contributed to addressing the core research questions and addressing the research objectives of this thesis. In addition to this, a critical reflection of the main implications and limitations of this research is provided. Finally, this thesis concludes with a discussion on potential future research directions to extend or further improve the work

8.1 Conclusions

The past few decades have established how digital technologies and platforms have provided an effective medium for spreading hateful content, bringing new challenges for behaviour-monitoring agencies [1]. As a result of the catastrophic impacts caused by online hate, research into this phenomenon is essential to understand it and negate its effects. Academic research has presented online hate as a complex phenomenon, with its definition evolving across several theoretical paradigms and disciplines. This has led to research within this field being fragmented throughout numerous disciplines, including computational social science (CSS), which have

introduced their own approaches to analysing associated issues. Despite all these extensive approaches and methods proposed to analyse online hate, limited research has explored how hateful behaviours and content compare across different online platforms. This thesis aims to address these limitations by proposing a cross-platform analysis framework to the study of online hate.

To ensure that this framework would address research limitations, the current research landscape in the field of online hate was first critically reviewed to identify areas where further insight was necessary — this literature review is provided in Chapter 2. Though it became apparent in this review that there are various outstanding issues in tackling online hate or reducing its impact, it was particularly evident that an enhanced approach or methodology would be required to gain a more comprehensive understanding. More specifically, previous research in online hate generally focussed around one particular platform, even with sufficient evidence showing that hate groups often strategize the usage of different online platforms in order to circumvent current monitoring efforts. This has resulted in a restricted, and at times unrealistic, understanding of this field. Cross-platform analysis would therefore be an effective approach to address this gap by advancing and validating existing findings on online hate.

After identifying gaps within the literature, in order to ensure that the analysis framework developed in this thesis would meet the requirements of online-hate researchers, existing analysis frameworks for online hate proposed in previous works were compared. This was carried out with particular regard to the functionalities they offered, the analysis methods they used, and the platforms they were concerned with. This comparison was also used to understand the extent to which existing analysis frameworks were applicable to hateful content on multiple online platforms, the findings from which are provided in Chapter 2.

Through this, it was found that, despite the extensive number of frameworks proposed, very few frameworks were designed to explore how hateful behaviours and content compare and relate across different online platforms. Existing research applying this approach for analysis was shown to be very limited, generally providing

preliminary insights. The insights gained from this literature review were used to answer the first research question — ***RQ1:*** *What are the main features and functionalities of current online-hate analysis frameworks and tools, and how do they deal with emerging research themes?*

The findings gained from answering the previous research question then informed the development of a novel framework to aid CSS researchers in gaining a more realistic image of the online global hate ecology. This could then inform how platform providers or law-enforcement agencies can potentially diminish its impacts, online and, in turn, offline. Chapter 4 details the designing process and structure of the cross-platform analysis framework. As well as providing novel perspectives into the usage of multiple platforms in online hate, this framework was developed to ensure that combined methods such as content and sentiment analysis, social network analysis, and content-diffusion analysis were prioritised in the framework functionalities, so as to address current gaps within academic literature. The development of this framework also answered the second research question of this thesis — ***RQ2:*** *How can existing online-hate analysis frameworks be further improved in order to enhance analysis efforts of online-hate researchers and to address major gaps in the current research landscape?*

In order to demonstrate how the cross-platform analysis framework could be used in online hate research in practice, the framework was applied to two primary case studies to investigate various hypotheses. Chapter 5 details the application of the framework to hateful content collected from the 2020 US election and the COVID-19 pandemic to gain a comprehensive understanding of how hateful users adapt and modify their content and behaviours across different online platforms. In particular, this analysis included harnessing various NLP and sentiment analysis techniques to compare the posting frequency, major topics of discussion, general sentiment and the overall linguistic composition of posts collected from Twitter, Reddit, 4chan and Stormfront.

Some of the key insights highlighted by the cross-platform analysis framework include the observation that offline contemporary events have more of an impact on

online activity on mainstream platforms rather than smaller underground platforms, such as Stormfront. Another observation made here is the fact that the more mainstream platforms like Twitter and Reddit would often take advantage of their large audience size to encourage other users to partake in certain actions, like joining protests and rallies or registering to vote. As well as this, linguistic analysis of both case studies showed that the higher usage of plural third-person pronouns reflected a stronger sense of the "Us Vs. Them" dichotomy in underground platforms like 4chan and Stormfront, which often promoted hateful narratives regarding groups of "others".

The framework was next applied to the study of how content is shared and diffused across multiple platforms, as detailed in Chapter 6. More specifically, this analysis involved the usage of various SNA techniques to create and explore networks of shared content, namely through the sharing of URL domains, across the election and COVID-19-related datasets from all four platforms. From this analysis, it became apparent that URLs played a huge role in content-sharing practices on all of the platforms to promote and support certain narratives, including campaigns like the 'Stop the Steal' movement, and false conspiracies regarding the cause of the COVID-19 pandemic. Such behaviour was most commonly found on mainstream platforms like Twitter, but significantly less on more underground platforms like Stormfront. Through analysing network graphs of shared domains across all four platforms, it was found that users from Reddit and 4chan would frequently post similar types of content from the same domains, thus implying a higher diffusion rate between both these platforms in particular.

Finally, this research reflected on the validity and effectiveness of the cross-platform analysis framework in investigating hateful content through the use of various validation criteria. This validation process highlighted the strengths of the analysis framework, particularly in terms of the unique perspectives it was able to provide in the study of the content and behaviours in online hate. Overall, it was found that adopting a cross-platform approach to the research of online hate provided a more comprehensive understanding of the issue as compared to a single-platform

approach, as well as a more extensive representation of the wider hate ecosystem. In addition to outlining its strengths, the validation process also indicated certain limitations of the framework, which are discussed in the following section. The insights obtained from the application of the framework to various case studies, as well as the validation process of the framework, answered the third and final research question of this thesis — ***RQ3:*** *What insights and understanding can be derived by analysing hateful content and activities across multiple online platforms?*

8.1.1 Practical Applications and Contributions

In addition to using the cross-platform analysis framework to address the research questions of this thesis, the proposed framework can be adapted and used in various disciplines outside of online hate, as well as practical applications beyond academia. This would involve understanding the underlying principles of the framework and tailoring its methodologies to suit specific contexts. For instance, such a framework could be used to analyse political discourse in order to identify patterns of polarisation and misinformation. This would provide insights into the dynamics of public opinion, policy debates, and potential impacts on democratic processes. Similarly, the framework could be utilised to analyse the tone and sentiment of news articles and comments to provide insight into media bias, public reception, and the spread of misinformation.

Another practical application of this cross-platform analysis framework is in market research. Here, the framework could be utilised to analyse consumer feedback and reviews across different e-commerce platforms, or social-media conversations related to products and services. This would enable businesses to understand consumer sentiment, identify emerging trends and adjust marketing strategies accordingly. This framework could further be adapted to study online discourse related to social movements and activism, identifying influential narratives, support networks, and potential strategies for amplifying causes. A cross-platform approach to analysis could also be used to study the linguistic and cultural aspects of online

communication across different platforms, providing insights into how language and identity are expressed and negotiated in the digital sphere.

In each of the use cases discussed here, the core methodologies utilised in the framework — including text analysis, classification, and statistical insights — are adapted to various contexts to extract meaningful information and patterns. By leveraging these methodologies and the proposed cross-platform approach, researchers and practitioners alike can gain valuable insights into a wide range of disciplines and contribute to evidence-based decision-making and interventions.

The versatility of the proposed cross-platform analysis framework in terms of its practical applications highlights the impacts and novelty that it offers in fostering multidisciplinary collaboration and informed decision-making. As demonstrated by its application to the two case studies detailed in this thesis, this framework encourages collaborations amongst researchers and experts from computer science, linguistics, sociology, psychology and data science, alongside other disciplines. Such collaboration enriches research by offering unique perspectives, which can lead to more comprehensive solutions. Additionally, by including techniques that can be applied across various platforms and contexts, this framework enables researchers to apply their findings from one domain to another in a more seamless manner. This promotes knowledge sharing and facilitates the leveraging of insights gained in one field to another.

This cross-platform analysis framework can also be used to foster discussions about ethical considerations in online hate or social-media research. This includes issues related to privacy, bias and freedom of speech. This would in turn increase efforts from researchers to develop strategies for mitigating biases and protecting individuals' rights when such research is conducted. Additionally, such a framework could be applied to the study of online activity over time, allowing researchers to monitor how online behaviour evolves and adapts across various contexts. This longitudinal analysis could provide insights into emerging trends allowing practitioners and policy-makers to provide more dynamic solutions. Similarly, this cross-platform analysis framework would further provide evidence-based insights for

policy development and enforcement. Policy-makers would therefore have access to a broader range of expertise and data-driven recommendations.

8.2 Limitations and Implications

While the cross-platform analysis framework developed in this thesis has demonstrated promising results in providing novel insights in the study of online hate, there are certain limitations and implications within the research that should be taken into consideration.

Firstly, although the applicability of the framework was validated through the use of multiple case studies, the scope of this thesis meant that the analysis was conducted using data from only three contemporary social events — the 2020 US election, the COVID-19 pandemic, and the 2020 BLM protests. While these events provided a useful lens through which to explore the spread of online hate across multiple platforms, it is possible that the findings may not be representative of other events or contexts. Since the findings from this analysis of these case studies highlighted instances of behaviours that were specific to certain contexts, for instance, a significantly higher usage of URLs on Reddit during the election as compared to during the pandemic, the framework would need to be applied to several other case studies in order to refine its design.

Additionally, the case studies that were used to assess the applicability of the cross-platform analysis framework were all previously reported to trigger hateful narratives and discourse amongst online users. Thus, it is unclear whether this may have impacted the performance of the framework on data collected during times outside of such events. Another limitation is that the framework was developed specifically to analyse online hate related to white supremacist and far-right hate ideologies. While this is certainly an important and pervasive form of online hate that has been linked to several catastrophic extremist attacks, it is important to acknowledge that there are many other forms of online hate that may require a different approach, or further adjustment of the framework.

Furthermore, the analysis was conducted using data from only four platforms — Twitter, Reddit, 4chan, and Stormfront. While these platforms each provide a different type of online environment, including both mainstream and underground platforms, it is possible that the framework may not produce similar insights when applied to other platforms or online spaces. It is important to note that different platforms may have unique characteristics and user demographics, which may affect the spread of hateful content and types of hateful behaviours. Similarly, since the cross-platform framework provides a general approach for analysis that can be applied to all platforms simultaneously, it may not always include specific platform features and functionalities, such as the usage of hashtags for engagement on Twitter.

Finally, while the framework was developed based on research limitations identified through an extensive review of academic literature and validated through the analysis of multiple case studies, it would be important to evaluate the framework with other researchers, particularly those from CSS, to ensure its validity and reliability. This would provide much needed insight into the usability of the framework and the insights that can be obtained with it, and would provide the necessary feedback required to refine the design of the framework further.

8.3 Future Work

Developing a cross-platform analysis framework for online hate research is a critical first step in providing a more comprehensive understanding of the wider networks of hateful content in online hate. However, there is still much work to be done, and a range of potential research directions can be pursued to extend and enhance the findings gained from developing such a framework.

One potential avenue of research is to expand the data analysis to other case studies besides the ones examined within this thesis. This would provide a broader understanding of how online discourse surrounding contemporary events can influence the prevalence of hateful activity across platforms. For instance, the results from the US election case study could be compared with the findings gained from applying the framework to data collected from a previous election

to gain more concrete insight into hateful activity during elections, and would help identify trends in online hate over time.

Additionally, harnessing various theoretical frameworks to explore the impact of different cultural, social, and political factors on the prevalence of online hate within these case studies would also provide much needed insight into the wider behaviours of hateful users. The cross-platform analysis framework could be used to compare online hate across different countries and cultures, and to examine the role of different factors, such as social norms, political polarisation, and media coverage, in contributing to the problem of online hate.

Another potential research direction is to explore data from "normal" time frames, i.e. not specific to any particular event that could trigger or influence online discourse. This would provide a better understanding of the prevalence of online hate in everyday, online interactions. Furthermore, making use of data from other platforms beyond the four examined in the current study, as well as different types of content, such as images and videos, could expand the scope of the framework and make it more widely applicable.

In order to further assess the validity of the cross-platform analysis framework and the novel insights it can provide, the framework could be applied to use cases from previous academic studies. This would highlight the contributions that such a framework can provide and also emphasise how current research can be improved by using such an approach.

The framework could also be used to examine the relationship between online and offline hate speech and hate crimes. Researchers could analyse data from different platforms to identify patterns in online hate that may indicate an increased risk of offline hate crimes, or identify which platforms are most likely to influence individuals into carrying out hateful acts. This information can be used to inform law-enforcement efforts to prevent hate crimes, and to develop more effective strategies for addressing the root causes of online hate and hate crimes.

One major area of further work would be to create analysis tools for researchers, law enforcement, and practitioners based off of this cross-platform analysis framework, so as to bridge the skills gap needed for applying the framework in practice. This would involve developing user-friendly interfaces for the framework and making the analytical techniques accessible to those without advanced technical skills. This could aid in facilitating its use in a variety of settings and by a wider range of stakeholders.

Finally, the cross-platform approach developed in the current study could also be applied to enhance content moderation practices on platforms. This could involve using predictive models to anticipate when hateful content or narratives are likely to increase on certain platforms, allowing for proactive measures to be taken to mitigate the spread of online hate. Such a framework could be enhanced further to explore the impact of various policies and practices on the prevalence of online hate, so the effectiveness of these policies across different platforms could be further compared. This would be able to help identify which policies and practices are most effective in reducing the prevalence of hateful activity online.

Overall, there are many interesting directions for future research in this field, and the framework developed in this study could serve as a valuable tool for advancing the understanding of online hate and its impact on society.

References

[1] William Housley et al. "Membership categorisation and antagonistic Twitter formulations". In: *Discourse & Communication* 11.6 (2017), pp. 567–590. eprint: https://doi.org/10.1177/1750481317726932. URL: https://doi.org/10.1177/1750481317726932.

[2] Edgar Pacheco and Neil Melhuish. *Online Hate Speech: A Survey on Personal Experiences and Exposure Among Adult New Zealanders*. Tech. rep. SSRN, 2018.

[3] Sean M. Eddington. "The Communicative Constitution of Hate Organizations Online: A Semantic Network Analysis of "Make America Great Again"". In: *Social Media + Society* 4.3 (2018), p. 2056305118790763. eprint: https://doi.org/10.1177/2056305118790763. URL: https://doi.org/10.1177/2056305118790763.

[4] Michael Chau and Jennifer Xu. "Mining Communities and their Relationships in Blogs: A Study of Online Hate Groups". In: *International Journal of Human-Computer Studies* 65.1 (2007). Information security in the knowledge economy, pp. 57–70. URL: http://www.sciencedirect.com/science/article/pii/S1071581906001248.

[5] Iginio Gagliardone. "Mapping and Analysing Hate Speech Online". In: *SSRN Electronic Journal* (2015).

[6] Alexandra Olteanu et al. "The Effect of Extremist Violence on Hateful Speech Online". In: *In Proceedings of the International AAAI Conference on Web and Social Media* (2018).

[7] Imran Awan. "Islamophobia and Twitter: A Typology of Online Hate Against Muslims on Social Media". In: *Policy & Internet* 6.2 (2014), pp. 133–150. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1944-2866.POI364. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI364.

[8] Imran Awan. "Cyber-Extremism: ISIS and the Power of Social Media". In: *Society* 54.2 (2017), pp. 138–149. URL: https://doi.org/10.1007/s12115-017-0114-0.

[9] Mark Potok. *The Year in Hate and Extremism*. Tech. rep. Southern Poverty Law Center, 2017. URL: https://www.splcenter.org/fighting-hate/intelligence-report/2017/year-hate-and-extremism.

[10] Ann John et al. "Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review". In: *J Med Internet Res* 20.4 (Apr. 2018), e129. URL: https://doi.org/10.2196/jmir.9044.

[11] N. F. Johnson et al. "Hidden Resilience and Adaptive Dynamics of the Global Online Hate Ecology". In: *Nature* 573.7773 (2019), pp. 261–265. URL: https://doi.org/10.1038/s41586-019-1494-7.

[12] Fatima Zahrah, Jason R C Nurse, and Michael Goldsmith. "A Comparison of Online Hate on Reddit and 4chan: A Case study of the 2020 US Election". In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. ACM, 2022, pp. 1797–1800.

[13] Jason Baumgartner et al. "The Pushshift Reddit Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), pp. 830–839. URL: https://www.aaai.org/ojs/index.php/ICWSM/article/view/7347.

[14] Antonis Papasavva et al. "Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 885–894. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/7354.

[15] Matteo Zignani et al. "Mastodon Content Warnings: Inappropriate Contents in a Microblogging Platform". In: *Proceedings of the International AAAI Conference on Web and Social Media* 13.01 (July 2019), pp. 639–645. URL: https://www.aaai.org/ojs/index.php/ICWSM/article/view/3262.

[16] Swati Agarwal and Ashish Sureka. *Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website*. 2017. eprint: 1701.04931.

[17] Peter Burnap and Matthew Leighton Williams. "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making". In: *Proceedings of the Conference on the Internet, Policy & Politics*. 2014.

[18] Peter Burnap and Matthew Leighton Williams. "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making". In: *Policy & Internet* 7 (2015), pp. 223–242.

[19] Sai Saketh Aluru et al. *Deep Learning Models for Multilingual Hate Speech Detection*. 2020. eprint: 2004.06465.

[20] Matthew G Beatty. "The Power of Diffusion Networks: Learning to Detect the Spread of Hate Speech Online". PhD thesis. Harvard University, 2018. URL: http://nrs.harvard.edu/urn-3:HUL.InstRepos:38811533.

[21] Binny Mathew et al. "Spread of Hate Speech in Online Social Media". In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '19. Boston, Massachusetts, USA: Association for Computing Machinery, 2019, pp. 173–182. URL: https://doi.org/10.1145/3292522.3326034.

[22] Carsten Schwemmer. *Social Media Strategies of Right-Wing Movements – The Radicalization of Pegida*. Mar. 2018. URL: osf.io/ebw7m.

[23] Paula Fortuna and Sérgio Nunes. "A Survey on Automatic Detection of Hate Speech in Text". In: *ACM Computing Surveys* 51.4 (July 2018). URL: https://doi.org/10.1145/3232676.

[24] Matti Pohjonen. "Extreme Digital Speech". In: *Extreme Digital Speech: Contexts, Responses and Solutions* (2020).

[25] Nery Ramati. *The Legal Response of Western Democracies To Online Terrorism and its Impact on the Right to Privacy and Freedom of Expression*. Tech. rep. Vox-Pol, 2020.

[26] Shruti Phadke and Tanushree Mitra. "Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. URL: https://doi.org/10.1145/3313831.3376456.

[27] Rohit Kumar Chandaluri and Shruti Phadke. *Cross-Platform Data Collection and Analysis for Online Hate Groups*. Tech. rep. Virginia Tech, 2020.

[28] Vitaliy V. Kashpur et al. "Where Russian Online Nationalists Go When Their Communities are Banned: A Case Study of Russian Nationalism". In: *Nationalism and Ethnic Politics* 26.2 (2020), pp. 145–166. eprint: https://doi.org/10.1080/13537113.2020.1751921. URL: https://doi.org/10.1080/13537113.2020.1751921.

[29] Savvas Zannettou et al. "The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources". In: *Proceedings of the 2017 Internet Measurement Conference*. IMC '17. London, United Kingdom: Association for Computing Machinery, 2017, pp. 405–417. URL: https://doi.org/10.1145/3131365.3131390.

[30] Maura Conway et al. "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts". In: *Studies in Conflict & Terrorism* 42.1-2 (2019), pp. 141–160. eprint: https://doi.org/10.1080/1057610X.2018.1513984. URL: https://doi.org/10.1080/1057610X.2018.1513984.

[31] Gabriel Hine et al. "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web". In: *In Proceedings of the International AAAI Conference on Web and Social Media* (2017). URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670/14790.

[32] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The Spread of True and False News Online". In: *Science* 359.6380 (2018), pp. 1146–1151. eprint: https://science.sciencemag.org/content/359/6380/1146.full.pdf. URL: https://science.sciencemag.org/content/359/6380/1146.

[33] Liang Wu and Huan Liu. "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 637–645. URL: https://doi.org/10.1145/3159652.3159677.

[34] William J Brady et al. "Emotion Shapes the Diffusion of Moralized Content in Social Networks". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.28 (July 2017), pp. 7313–7318. URL: https://pubmed.ncbi.nlm.nih.gov/28652356%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5514704/.

[35] Björn Ross et al. *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*. 2017. arXiv: 1701.08118 [cs.CL].

[36] Zeerak Waseem. "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter". In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 138–142. URL: https://www.aclweb.org/anthology/W16-5618.

[37] European Union Commission. *Code of Conduct on Countering Illegal Hate Speech Online*. 2016.

[38] United Nations General Assembly. *International Convention on the Elimination of All Forms of Racial Discrimination*. 1965.

[39] ILGA. *Hate Crime and Hate Speech*. URL: https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech (visited on 06/14/2020).

[40] Chikashi Nobata et al. "Abusive Language Detection in Online User Content". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 145–153. URL: https://doi.org/10.1145/2872427.2883062.

[41] William Warner and Julia Hirschberg. "Detecting Hate Speech on the World Wide Web". In: *Proceedings of the Second Workshop on Language in Social Media*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 19–26. URL: https://www.aclweb.org/anthology/W12-2103.

[42] Thomas Davidson et al. *Automated Hate Speech Detection and the Problem of Offensive Language*. 2017.

[43] Joni Salminen et al. "Developing an online hate classifier for multiple social media platforms". In: *Human-Centric Computing and Information Sciences* 10.1 (2020), p. 1. URL: https://doi.org/10.1186/s13673-019-0205-6.

[44] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. "A Measurement Study of Hate Speech in Social Media". In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT '17. Prague, Czech Republic: Association for Computing Machinery, 2017, pp. 85–94. URL: https://doi.org/10.1145/3078714.3078723.

[45] Facebook. *Community Standards*. 2020. URL: https://transparency.fb.com/en-gb/policies/community-standards/.

[46] YouTube. *Hate Speech Policy*. 2020. URL: https://support.google.com/youtube/answer/2801939?hl=en-GB.

[47] Twitter. *Hateful Conduct Policy*. 2020. URL: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[48] Inari Sakki and Laura Castrén. "Dehumanization Through Humour and Conspiracies in Online Hate Towards Chinese People During the COVID-19 Pandemic". In: *British Journal of Social Psychology* 61.4 (Oct. 2022), pp. 1418–1438. URL: https://doi.org/10.1111/bjso.12543.

[49] Panos Kompatsiaris. “Whitewashing the Nation: Racist Jokes and the Construction of the African ‘Other’ in Greek Popular Cinema”. In: *Social Identities* 23.3 (2017), pp. 360–375. eprint: https://doi.org/10.1080/13504630.2016.1207513. URL: https://doi.org/10.1080/13504630.2016.1207513.

[50] MICHAEL BILLIG. “Humour and Hatred: The Racist Jokes of the Ku Klux Klan”. In: *Discourse & Society* 12.3 (2001), pp. 267–289. eprint: https://doi.org/10.1177/0957926501012003001. URL: https://doi.org/10.1177/0957926501012003001.

[51] Michael Billig. “Comic Racism and Violence”. In: *Beyond a Joke: The Limits of Humour*. London: Palgrave Macmillan UK, 2005, pp. 25–44. URL: https://doi.org/10.1057/9780230236776_2.

[52] Sara Douglass et al. “"They Were Just Making Jokes": Ethnic/Racial Teasing and Discrimination Among Adolescents.” In: *Cultural Diversity & Ethnic Minority Psychology* 22.1 (Jan. 2016), pp. 69–82.

[53] Irene Kwok and Yuzhou Wang. “Locate the Hate: Detecting Tweets against Blacks”. In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI’13. Bellevue, Washington: AAAI Press, 2013, pp. 1621–1622.

[54] Michal Ptaszynski et al. “In the Service of Online Order : Tackling Cyber-Bullying with Machine Learning and Affect Analysis”. In: *International Journal of Computational Linguistics Research* 1.3 (Sept. 2010), pp. 135–154. URL: http://hdl.handle.net/2115/63640.

[55] Ying Chen et al. “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. 2012, pp. 71–80.

[56] Simona Frenda et al. “Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter.” In: *Journal of Intelligent & Fuzzy Systems* 36.5 (2019), pp. 4743–4752.

[57] Karen M Douglas. “Psychology, Discrimination and Hate Groups Online”. In: *The Oxford Handbook of Internet Psychology* (2007), pp. 155–164.

[58] Brendesha M Tynes et al. “Virtual Environments, Online Racial Discrimination, and Adjustment Among a Diverse, School-based Sample of Adolescents”. In: *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)* 6.3 (2014), pp. 1–16.

[59] Neil Thompson. *Anti-Discriminatory Practice : Equality, Diversity and Social Justice*. English. London, 2016.

[60] Adi Kuntsman. “Belonging Through Violence: Flaming, Erasure, and Performativity in Queer Migrant Community”. In: *Queer online: Media technology and sexuality* (2007), pp. 101–120.

[61] Hangwoo Lee. “Behavioral Strategies for Dealing with Flaming in An Online Forum”. In: *The Sociological Quarterly* 46.2 (2005), pp. 385–403. eprint: https://doi.org/10.1111/j.1533-8525.2005.00017.x. URL: https://doi.org/10.1111/j.1533-8525.2005.00017.x.

[62] Wang, Kuei-Ing, and Jou-Fan Shih. "Factors Influencing University Students' Online Disinhibition Behavior – The Moderating Effects of Deterrence and Social Identity". In: *International Journal of Economics and Management Engineering* 8.5 (2014), pp. 1486–1492. URL: https://publications.waset.org/vol/89.

[63] Jungyong Lee and Changhyun Jin. "The Relationship Between Self-Concepts and Flaming Behaviour: Polarity of The Online Comments". In: *Journal of Theoretical and Applied Information Technology* 97.19 (2019).

[64] Matthew Costello, James Hawdon, and Thomas N. Ratliff. "Confronting Online Extremism: The Effect of Self-Help, Collective Efficacy, and Guardianship on Being a Target for Hate Speech". In: *Social Science Computer Review* 35.5 (2017), pp. 587–605. eprint: https://doi.org/10.1177/0894439316666272. URL: https://doi.org/10.1177/0894439316666272.

[65] Matthew Costello et al. "Predictors of Viewing Online Extremism Among America's Youth". In: *Youth & Society* 52.5 (2020), pp. 710–727. eprint: https://doi.org/10.1177/0044118X18768115. URL: https://doi.org/10.1177/0044118X18768115.

[66] Lacy G. McNamee, Brittany L. Peterson, and Jorge Peña. "A Call to Educate, Participate, Invoke and Indict: Understanding the Communication of Online Hate Groups". In: *Communication Monographs* 77.2 (2010), pp. 257–280. eprint: https://doi.org/10.1080/03637751003758227. URL: https://doi.org/10.1080/03637751003758227.

[67] Sheryl Prentice et al. "Analyzing the Semantic Content and Persuasive Composition of Extremist Media: A Case Study of Texts Produced During the Gaza Conflict". In: *Information Systems Frontiers* 13.1 (2011), pp. 61–73. URL: https://doi.org/10.1007/s10796-010-9272-y.

[68] Daniele Conversi. "Irresponsible Radicalisation: Diasporas, Globalisation and Long-Distance Nationalism in the Digital Age". In: *Journal of Ethnic and Migration Studies* 38.9 (2012), pp. 1357–1379. eprint: https://doi.org/10.1080/1369183X.2012.698204. URL: https://doi.org/10.1080/1369183X.2012.698204.

[69] A. Bermingham et al. "Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation". In: *2009 International Conference on Advances in Social Network Analysis and Mining*. 2009, pp. 231–236.

[70] Jonathan Greenblatt. *When Hateful Speech Leads to Hate Crimes: Taking Bigotry Out of the Immigration Debate*. 2015. URL: https://www.adl.org/blog/when-hateful-speech-leads-to-hate-crimes-taking-bigotry-out-of-the-immigration-debate (visited on 07/07/2020).

[71] Olga Jubany and Malin Roiha. *Backgrounds, Experiences and Responses to Online Hate Speech: A Comparative Cross-Country Analysis*. Tech. rep. The Prism Project, 2015.

[72] Nathan Jurgenson. "Digital Dualism and the Fallacy of Web Objectivity". In: *The Society Pages* 2011 (2011).

[73] Mirca Madianou and Daniel Miller. "Polymedia: Towards a New Theory of Digital Media in Interpersonal Communication". In: *International Journal of Cultural Studies* 16.2 (2013), pp. 169–187. eprint: https://doi.org/10.1177/1367877912452486. URL: https://doi.org/10.1177/1367877912452486.

[74] David Sullivan. *The Consequences of Legislating Cyberlaw After Terrorist Attacks*. 2019. URL: https://www.justsecurity.org/63560/the-consequences-of-legislating-cyberlaw-after-terrorist-attacks/ (visited on 07/27/2020).

[75] Mubaraz Ahmed. "Impact of Content". In: *Extreme Digital Speech: Contexts, Responses and Solutions* (2020).

[76] Deen Freelon. "On the Interpretation of Digital Trace Data in Communication and Social Computing Research". In: *Journal of Broadcasting & Electronic Media* 58.1 (2014), pp. 59–75. URL: https://doi.org/10.1080/08838151.2013.875018.

[77] Scott A. Golder and Michael W. Macy. "Digital Footprints: Opportunities and Challenges for Online Social Research". In: *Annual Review of Sociology* 40.1 (2014), pp. 129–152. URL: https://doi.org/10.1146/annurev-soc-071913-043145.

[78] Keith N. Hampton. "Studying the Digital: Directions and Challenges for Digital Methods". In: *Annual Review of Sociology* 43.1 (2017), pp. 167–188. URL: https://doi.org/10.1146/annurev-soc-060116-053505.

[79] David Lazer and Jason Radford. "Data ex Machina: Introduction to Big Data". In: *Annual Review of Sociology* 43.1 (2017), pp. 19–39. URL: https://doi.org/10.1146/annurev-soc-060116-053457.

[80] Shawn Walker, Dan Mercea, and Marco Bastos. "The Disinformation Landscape and the Lockdown of Social Platforms". In: *Information, Communication & Society* 22.11 (2019), pp. 1531–1543. URL: https://doi.org/10.1080/1369118X.2019.1648536.

[81] Deen Freelon. "Computational Research in the Post-API Age". In: *Political Communication* 35.4 (2018), pp. 665–668. URL: https://doi.org/10.1080/10584609.2018.1477506.

[82] Cornelius Puschmann. "An end to the Wild West of Social Media Research: A Response to Axel Bruns". In: *Information, Communication & Society* 22.11 (2019), pp. 1582–1589. URL: https://doi.org/10.1080/1369118X.2019.1646300.

[83] Danah Boyd and Kate Crawford. "Critical Questions for Big Data". In: *Information, Communication & Society* 15.5 (2012), pp. 662–679. URL: https://doi.org/10.1080/1369118X.2012.678878.

[84] Lev Manovich. "Trending: The Promises and the Challenges of Big Social Data". eng. In: *Debates in the Digital Humanities*. University of Minnesota Press, 2012. Chap. 27. URL: https://minnesota.universitypressscholarship.com/10.5749/minnesota/9780816677948.001.0001/upso-9780816677948-chapter-47.

[85] Alexander Halavais. "Overcoming Terms of Service: A Proposal for Ethical Distributed Research". In: *Information, Communication & Society* 22.11 (2019), pp. 1567–1581. URL: https://doi.org/10.1080/1369118X.2019.1627386.

[86] Komal Patel. "Testing the Limits of the First Amendment: How a CFAA Prohibition on Online Antidiscrimination Testing Infringes on Protected Speech Activity". In: *Columbia Law Review* 118.5 (2017). URL: https://dx.doi.org/10.2139/ssrn.3046847.

[87] Jason Baumgartner et al. "The Pushshift Telegram Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 840–847. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/7348.

[88] Emilija Jokubauskaite and Stijn Peeters. "Generally Curious: Thematically Distinct Datasets of General Threads on 4chan/pol/". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 863–867. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/7351.

[89] Gabriel Fair and Ryan Wesslen. "Shouting into the Void: A Database of the Alternative Social Media Platform Gab". In: *Proceedings of the International AAAI Conference on Web and Social Media* 13.01 (July 2019), pp. 608–610. URL: https://www.aaai.org/ojs/index.php/ICWSM/article/view/3258.

[90] Antigoni Founta et al. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior". In: *Proceedings of the International AAAI Conference on Web and Social Media* (2018). URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909.

[91] Savvas Zannettou et al. "What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber". In: *Companion Proceedings of the The Web Conference 2018* (2018).

[92] Val Burris, Emery Smith, and Ann Strahm. "White Supremacist Networks on the Internet". In: *Sociological Focus* 33.2 (2000), pp. 215–235. eprint: https://doi.org/10.1080/00380237.2000.10571166. URL: https://doi.org/10.1080/00380237.2000.10571166.

[93] Emre Calisir and Marco Brambilla. "The Long-Running Debate about Brexit on Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 848–852. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/7349.

[94] Giovanni Brena et al. "News Sharing User Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions". In: *Proceedings of the International AAAI Conference on Web and Social Media* 13.01 (July 2019), pp. 592–597. URL: https://www.aaai.org/ojs/index.php/ICWSM/article/view/3256.

[95] Venkata Rama Kiran Garimella and Gareth Tyson. "WhatsApp, Doc? A First Look at WhatsApp Public Group Data". In: *ArXiv* abs/1804.01473 (2018).

[96] Monika Fishman and Brian Bickert. *Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?* 2017. URL: https://about.fb.com/news/2018/04/keeping-terrorists-off-facebook/ (visited on 06/01/2020).

[97] Susan Wojcicki. *Expanding Our Work Against Abuse of Our Platform*. 2017. (Visited on 06/01/2020).

[98] Twitter. *Transparency Report: Twitter Rules Enforcement*. 2019. (Visited on 06/02/2020).

[99] Stuart Macdonald. *How Tech Companies are Successfully Disrupting Terrorist Social Media Activity*. 2018. URL: https://theconversation.com/how-tech-companies-are-successfully-disrupting-terrorist-social-media-activity-98594 (visited on 06/01/2020).

[100] Marcos Zampieri et al. *Predicting the Type and Target of Offensive Posts in Social Media*. 2019. eprint: 1902.09666.

[101] Stéphan Tulkens et al. *A Dictionary-based Approach to Racism Detection in Dutch Social Media*. 2016. eprint: 1608.08738.

[102] Fabio Del Vigna et al. "Hate Me, Hate Me Not: Hate Speech Detection on Facebook". In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)* (2017).

[103] K. Nugroho et al. "Improving Random Forest Method to Detect Hatespeech and Offensive Word". In: *2019 International Conference on Information and Communications Technology (ICOIACT)*. 2019, pp. 514–518.

[104] Joni Salminen et al. "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media". In: *Twelfth International AAAI Conference on Web and Social Media*. 2017.

[105] Pinkesh Badjatiya et al. "Deep Learning for Hate Speech Detection in Tweets". In: *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion* (2017). URL: http://dx.doi.org/10.1145/3041021.3054223.

[106] Pete Burnap and Matthew L Williams. "Us and Them: Identifying Cyber Hate on Twitter Across Multiple Protected Characteristics". In: *EPJ Data Science* 5.1 (2016), p. 11. URL: https://doi.org/10.1140/epjds/s13688-016-0072-6.

[107] Abid Hussain Wani, Nahida Shafi Molvi, and Sheikh Ishrah Ashraf. "Detection of Hate and Offensive Speech in Text". In: *Intelligent Human Computer Interaction*. Ed. by Uma Shanker Tiwary and Santanu Chaudhury. Cham: Springer International Publishing, 2020, pp. 87–93.

[108] Sattam Almatarneh et al. "Supervised Classifiers to Identify Hate Speech on English and Spanish Tweets". In: *Digital Libraries at the Crossroads of Digital Information for the Future*. Ed. by Adam Jatowt, Akira Maeda, and Sue Yeon Syn. Cham: Springer International Publishing, 2019, pp. 23–30.

[109] P. S. Br Ginting, B. Irawan, and C. Setianingsih. "Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method". In: *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. 2019, pp. 105–111.

[110] O. Oriola and E. Kotzé. "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets". In: *IEEE Access* 8 (2020), pp. 21496–21509.

[111] Nemanja Djuric et al. “Hate Speech Detection with Comment Embeddings”. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW ’15 Companion. Florence, Italy: Association for Computing Machinery, 2015, pp. 29–30. URL: https://doi.org/10.1145/2740908.2742760.

[112] Zeerak Waseem and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”. In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93. URL: https://www.aclweb.org/anthology/N16-2013.

[113] Jae Yeon Kim et al. *Intersectional Bias in Hate Speech and Abusive Language Datasets*. 2020. arXiv: 2005.05921 [cs.CL].

[114] T. Lynn et al. “A Comparison of Machine Learning Approaches for Detecting Misogynistic Speech in Urban Dictionary”. In: *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*. 2019, pp. 1–8.

[115] S. Ahammed et al. “Implementation of Machine Learning to Detect Hate Speech in Bangla Language”. In: *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. 2019, pp. 317–320.

[116] Shuhua Liu and Thomas Forss. “Combining N-Gram Based Similarity Analysis with Sentiment Analysis in Web Content Classification”. In: *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1*. IC3K 2014. Rome, Italy: SCITEPRESS - Science and Technology Publications, Lda, 2014, pp. 530–537. URL: https://doi.org/10.5220/0005170305300537.

[117] Shuhan Yuan, Xintao Wu, and Yang Xiang. “A Two Phase Deep Learning Model for Identifying Discrimination from Tweets”. In: *Proceedings of 19th International Conference on Extending Database Technology (EDBT)*. 2016.

[118] S. Jaki et al. “Online Hatred of Women in the Incels.me Forum: Linguistic Analysis and Automatic Detection”. In: *Journal of Language Aggression and Conflict* 7.2 (2019), pp. 240–268.

[119] Zewdie Mossie and Jenq-Haur Wang. “Vulnerable Community Identification Using Hate Speech Detection on Social Media”. In: *Information Processing & Management* 57.3 (2020), p. 102087. URL: https://www.sciencedirect.com/science/article/pii/S0306457318310902.

[120] Prashant Kapil, Asif Ekbal, and Dipankar Das. *Investigating Deep Learning Approaches for Hate Speech Detection in Social Media*. 2020. arXiv: 2005.14690 [cs.CL].

[121] S. Liu and T. Forss. “New Classification Models for Detecting Hate and Violence Web Content”. In: *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*. Vol. 1. 2015, pp. 487–495.

[122] Maral Dadvar et al. “Improved Cyberbullying Detection Using Gender Information”. In: *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. 2012.

[123] Karthik Dinakar, Roi Reichart, and Henry Lieberman. "Modeling the Detection of Textual Cyberbullying". In: *The Social Mobile Web*. 2011.

[124] Edel Greevy and Alan F. Smeaton. "Classifying Racist Texts Using a Support Vector Machine". In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, United Kingdom: Association for Computing Machinery, 2004, pp. 468–469. URL: https://doi.org/10.1145/1008992.1009074.

[125] Yashar Mehdad and Joel Tetreault. "Do Characters Abuse More Than Words?" In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles: Association for Computational Linguistics, Sept. 2016, pp. 299–303. URL: https://www.aclweb.org/anthology/W16-3638.

[126] Anna Schmidt and Michael Wiegand. "A Survey on Hate Speech Detection using Natural Language Processing". In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. URL: https://www.aclweb.org/anthology/W17-1101.

[127] Njagi Dennis Gitari et al. "A Lexicon-based Approach for Hate Speech Detection". In: *International Journal of Multimedia and Ubiquitous Engineering* 10 (2015), pp. 215–230.

[128] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[129] Keith Cortis and Siegfried Handschuh. "Analysis of Cyberbullying Tweets in Trending World Events". In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business*. i-KNOW '15. Graz, Austria: Association for Computing Machinery, 2015. URL: https://doi.org/10.1145/2809563.2809605.

[130] Fatima Zahrah, Jason RC Nurse, and Michael Goldsmith. "#ISIS vs #ActionCountersTerrorism: A Computational Analysis of Extremist and Counter-extremist Twitter Narratives". In: *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. 2020, pp. 438–447.

[131] David Robinson, Ziqi Zhang, and Jonathan Tepper. "Hate Speech Detection on Twitter: Feature Engineering v.s. Feature Selection". In: *The Semantic Web: ESWC 2018 Satellite Events*. Cham: Springer International Publishing, 2018, pp. 46–49.

[132] Valentino Santucci et al. "Detecting Hate Speech for Italian Language in Social Media". In: *EVALITA 2018, co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*. Vol. 2263. 2018.

[133] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades". In: *Journal of Economic Perspectives* 12.3 (Sept. 1998), pp. 151–170. URL: https://www.aeaweb.org/articles?id=10.1257/jep.12.3.151.

[134] Everett M Rogers. *Diffusion of Innovations*. eng. 5th ed. Ebook central. Riverside: Free Press, 2003.

[135] Jure Leskovec, Lada Adamic, and Bernardo Huberman. "The Dynamics of Viral Marketing". eng. In: *ACM Transactions on the Web (TWEB)* 1.1 (2007), 5–es.

[136] Norman T. J Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. eng. 2nd ed. London: Griffin, 1975.

[137] Frank M. Bass. "A New Product Growth for Model Consumer Durables". In: *Management Science* 15.5 (1969), pp. 215–227. URL: http://www.jstor.org/stable/2628128.

[138] Thomas Valente. "Network Models of the Diffusion of Innovations". eng. In: *Computational & Mathematical Organization Theory* 2.2 (1996), pp. 163–164.

[139] Wendi Wang et al. "Mathematical Models of Innovation Diffusion with Stage Structure". In: *Applied Mathematical Modelling* 30.1 (2006), pp. 129–146. URL: http://www.sciencedirect.com/science/article/pii/S0307904X05000533.

[140] Mark Granovetter. "Threshold Models of Collective Behavior". In: *American Journal of Sociology* 83.6 (1978), pp. 1420–1443. URL: http://www.jstor.org/stable/2778111.

[141] Stephen Morris. "Contagion". In: *The Review of Economic Studies* 67.1 (2000), pp. 57–78.

[142] P.S. Dodds and D.J. Watts. "Universal Behavior in a Generalized Model of Contagion". English. In: *Physical Review Letters* 92.21 (2004), pp. 218701–218701.

[143] M. L. Markus. "Toward a "Critical Mass" Theory of Interactive Media: Universal Access, Interdependence and Diffusion". In: *Communication Research* 14.5 (1987), pp. 491–511. eprint: https://doi.org/10.1177/009365087014005003. URL: https://doi.org/10.1177/009365087014005003.

[144] Rabikar Chatterjee and Jehoshua Eliashberg. "The Innovation Diffusion Process in a Heterogeneous Population: A Micromodeling Approach". In: *Management Science* 36.9 (1990), pp. 1057–1079. URL: http://www.jstor.org/stable/2632356.

[145] Michael Kearns, Siddharth Suri, and Nick Montfort. "An Experimental Study of the Coloring Problem on Human Subject Networks". In: *Science* 313.5788 (2006), pp. 824–827. eprint: https://science.sciencemag.org/content/313/5788/824.full.pdf. URL: https://science.sciencemag.org/content/313/5788/824.

[146] Michael Kearns et al. "Behavioral Experiments on Biased Voting in Networks". In: *Proceedings of the National Academy of Sciences* 106.5 (2009), pp. 1347–1352. eprint: https://www.pnas.org/content/106/5/1347.full.pdf. URL: https://www.pnas.org/content/106/5/1347.

[147] Florian Buhl, Elisabeth Günther, and Thorsten Quandt. "Observing the Dynamics of the Online News Ecosystem". In: *Journalism Studies* 19.1 (2018), pp. 79–104. eprint: https://doi.org/10.1080/1461670X.2016.1168711. URL: https://doi.org/10.1080/1461670X.2016.1168711.

[148] Lars Backstrom et al. "Group Formation in Large Social Networks: Membership, Growth, and Evolution". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 44–54. URL: https://doi.org/10.1145/1150402.1150412.

[149] Kyungmo Kim, Young Min Baek, and Narae Kim. "Online News Diffusion Dynamics and Public Opinion Formation: A Case Study of the Controversy Over Judges' Personal Opinion Expression on SNS in Korea". In: *The Social Science Journal* 52.2 (2015), pp. 205–216. URL: http://www.sciencedirect.com/science/article/pii/S0362331915000269.

[150] Eni Mustafaraj and Panagiotis Takis Metaxas. "The Fake News Spreading Plague: Was It Preventable?" In: *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17. Troy, New York, USA: Association for Computing Machinery, 2017, pp. 235–239. URL: https://doi.org/10.1145/3091478.3091523.

[151] Justin Cheng et al. "Can Cascades Be Predicted?" In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. Seoul, Korea: Association for Computing Machinery, 2014, pp. 925–936. URL: https://doi.org/10.1145/2566486.2567997.

[152] Justin Cheng et al. "Do Cascades Recur?" In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 671–681. URL: https://doi.org/10.1145/2872427.2882993.

[153] Sharad Goel et al. "The Structural Virality of Online Diffusion". In: *Manag. Sci.* 62 (2016), pp. 180–196.

[154] Michela Del Vicario et al. "The Spreading of Misinformation Online". In: *Proceedings of the National Academy of Sciences* 113.3 (2016), pp. 554–559. eprint: https://www.pnas.org/content/113/3/554.full.pdf. URL: https://www.pnas.org/content/113/3/554.

[155] Adrien Friggeri et al. "Rumor Cascades". In: *Proceedings of the Eighth International Conference on Web and Social Media*. Ed. by Eytan Adar et al. The AAAI Press, 2014. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122.

[156] Fang Jin et al. "Epidemiological Modeling of News and Rumors on Twitter". In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. SNAKDD '13. Chicago, Illinois: Association for Computing Machinery, 2013. URL: https://doi.org/10.1145/2501025.2501027.

[157] Jure Leskovec et al. "Patterns of Cascading Behavior in Large Blog Graphs". In: *Proceedings of the 2007 SIAM International Conference on Data Mining*. 2007, pp. 551–556. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611972771.60. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611972771.60.

[158] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 1395–1405. URL: https://doi.org/10.1145/2736277.2741637.

[159] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. "Virality Prediction and Community Structure in Social Networks". In: *Scientific Reports* 3.1 (2013), p. 2522. URL: https://doi.org/10.1038/srep02522.

[160] C. Remy et al. "Information Diffusion on Twitter: Everyone Has Its Chance, But All Chances Are Not Equal". In: *2013 International Conference on Signal-Image Technology Internet-Based Systems*. 2013, pp. 483–490.

[161] Daniel Preoţiuc-Pietro et al. "Beyond Binary Labels: Political Ideology Prediction of Twitter Users". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 729–740. URL: https://www.aclweb.org/anthology/P17-1068.

[162] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. "Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. Hyderabad, India: Association for Computing Machinery, 2011, pp. 695–704. URL: https://doi.org/10.1145/1963405.1963503.

[163] Shaomei Wu et al. "Does Bad News Go Away Faster?" In: *ICWSM*. 2011.

[164] Teo Keipi et al. *Online Hate and Harmful Content: Cross-National Perspectives*. Routledge, Dec. 2016.

[165] Maryann M. Durland and Kimberly A. Fredericks. "An Introduction to Social Network Analysis". In: *New Directions for Evaluation* 2005.107 (2005), pp. 5–13. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ev.157. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ev.157.

[166] David A. Snow, Louis A. Zurcher, and Sheldon Ekland-Olson. "Social Networks and Social Movements: A Microstructural Approach to Differential Recruitment". In: *American Sociological Review* 45.5 (1980), pp. 787–801. URL: http://www.jstor.org/stable/2094895.

[167] Naomi Rosenthal et al. "Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State". In: *American Journal of Sociology* 90.5 (1985), pp. 1022–1054. URL: http://www.jstor.org/stable/2780088.

[168] Doug McAdam. "Recruitment to High-Risk Activism: The Case of Freedom Summer". In: *American Journal of Sociology* 92.1 (1986), pp. 64–90. URL: http://www.jstor.org/stable/2779717.

[169] Roberto M. Fernandez and Doug McAdam. "Social Networks and Social Movements: Multiorganizational Fields and Recruitment to Mississippi Freedom Summer". In: *Sociological Forum* 3.3 (1988), pp. 357–382. URL: http://www.jstor.org/stable/684338.

[170] Gerald Marwell, Pamela E. Oliver, and Ralph Prahl. "Social Networks and Collective Action: A Theory of the Critical Mass. III". In: *American Journal of Sociology* 94.3 (1988), pp. 502–534. URL: http://www.jstor.org/stable/2780252.

[171] Roger V. Gould. "Multiple Networks and Mobilization in the Paris Commune, 1871". In: *American Sociological Review* 56.6 (1991), pp. 716–729. URL: http://www.jstor.org/stable/2096251.

[172] Roger V. Gould. "Collective Action and Network Structure". In: *American Sociological Review* 58.2 (1993), pp. 182–196. URL: http://www.jstor.org/stable/2095965.

[173] Roger V. Gould. "Patron-Client Ties, State Centralization, and the Whiskey Rebellion". In: *American Journal of Sociology* 102.2 (1996), pp. 400–429. URL: http://www.jstor.org/stable/2782630.

[174] Hyojoung Kim and Peter S. Bearman. "The Structure and Dynamics of Movement Participation". In: *American Sociological Review* 62.1 (1997), pp. 70–93. URL: http://www.jstor.org/stable/2657453.

[175] Russell L. Curtis and Louis A. Zurcher. "Stable Resources of Protest Movements: The Multi-Organizational Field". In: *Social Forces* 52.1 (1973), pp. 53–61. URL: http://www.jstor.org/stable/2576423.

[176] John D. McCarthy and Mayer N. Zald. "Resource Mobilization and Social Movements: A Partial Theory". In: *American Journal of Sociology* 82.6 (1977), pp. 1212–1241. URL: http://www.jstor.org/stable/2777934.

[177] David Knoke. *Political Networks: The Structural Perspective*. Structural Analysis in the Social Sciences. Cambridge University Press, 1990.

[178] Y. Zhou et al. "US Domestic Extremist Groups on the Web: Link and Content Analysis". In: *IEEE Intelligent Systems* 20.5 (2005), pp. 44–51.

[179] Luca Tateo. "The Italian Extreme Right On-line Network: An Exploratory Study Using an Integrated Social Network Analysis and Content Analysis Approach". In: *Journal of Computer-Mediated Communication* 10.2 (2005), pp. 00–00. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1083-6101.2005.tb00247.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2005.tb00247.x.

[180] Courtney Weaver. *The Alt-Right is Revising its Online Strategy After a Backlash*. 2018. URL: https://www.ft.com/content/c94642da-9f08-11e8-85da-eeb7a9ce36e4.

[181] Fatima Zahrah and Jason R C Nurse. "Terrorism and Online Extremism". In: *Elgar Encyclopedia of Technology and Politics*. Edward Elgar Publishing, 2022, pp. 62–67.

[182] Leandro Araújo Silva et al. "Analyzing the Targets of Hate in Online Social Media". In: *International AAAI Conference on Web and Social Media*. 2016.

[183] Eshwar Chandrasekharan et al. "The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3175–3187. URL: https://doi.org/10.1145/3025453.3026018.

[184] Patricia Bazeley. *Qualitative Data Analysis: Practical Strategies*. SAGE Publications Inc., 2013.

[185] B Miles Matthew, Huberman A Michael, and Saldaña Johnny. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications Inc, 2014.

[186] J Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE Publications Inc, 2012. URL: https://books.google.co.uk/books?id=V3tTG4jvgFkC.

[187] M Schreier. *Qualitative Content Analysis in Practice*. SAGE Publications Inc, 2012. URL: https://books.google.co.uk/books?id=zTGhsdl17gYC.

[188] R V Kozinets. *Netnography: Redefined*. SAGE Publications Inc, 2015.

[189] Cornelius Puschmann and Jean Burgess. "The Politics of Twitter Data". In: *HIIG Discussion Paper Series* (2013). URL: http://dx.doi.org/10.2139/ssrn.2206225.

[190] Ahmed Waqas et al. "Mapping Online Hate: A Scientometric Analysis on Research Trends and Hotspots in Research on Online Hate". eng. In: *PloS one* 14.9 (Sept. 2019), e0222194–e0222194. URL: https://pubmed.ncbi.nlm.nih.gov/31557227%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6763199/.

[191] C Robson. *Real World Research: A Resource for Social Scientists and Practitioner-Researchers (3rd edition)*. Wiley, 2011.

[192] L Sloan and A Quan-Haase. *The SAGE Handbook of Social Media Research Methods*. SAGE Publications Inc, 2017.

[193] Jan H Kietzmann et al. "Social Media? Get serious! Understanding the Functional Building Blocks of Social Media". In: *Business Horizons* 54.3 (2011), pp. 241–251. URL: https://www.sciencedirect.com/science/article/pii/S0007681311000061.

[194] Joe F Hair, Michael Page, and Niek Brunsveld. *Essentials of Business Research Methods*. Routledge, 2019.

[195] Daniel Zeng et al. "Social Media Analytics and Intelligence". In: *IEEE Intelligent Systems* 25.6 (2010), pp. 13–16.

[196] Annette Markham and Elizabeth Buchanan. *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. Tech. rep. Association of Internet Researchers, 2012, p. 19. URL: https://aoir.org/reports/ethics2.pdf.

[197] V. Tablan et al. "GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud." In: *Philosophical Transactions of the Royal Society A* 371.1983 (2013). URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2012.0071.

[198] Joseph Mei and Richard Frank. "Sentiment Crawling: Extremist Content Collection Through a Sentiment Analysis Guided Web-Crawler". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ASONAM '15. Paris, France: Association for Computing Machinery, 2015, pp. 1024–1027.

[199] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. "Effective Hate-Speech Detection in Twitter Data Using Recurrent Neural Networks". In: *Applied Intelligence* 48.12 (2018), pp. 4730–4742. URL: https://doi.org/10.1007/s10489-018-1242-y.

[200] Lana Cuthbertson et al. "Women, Politics and Twitter: Using Machine Learning to Change the Discourse". In: *ArXiv* abs/1911.11025 (2019).

[201] Michele Corazza et al. "InriaFBK at Germeval 2018: Identifying Offensive Tweets Using Recurrent Neural Networks". In: *GermEval 2018 Workshop*. Vienna, Austria, Sept. 2018. URL: https://hal.archives-ouvertes.fr/hal-01906096.

[202] Joshua Uyheng and Kathleen M Carley. "Bots and Online Hate During the COVID-19 Pandemic: Case Studies in the United States and the Philippines". In: *Journal of Computational Social Science* 3.2 (2020), pp. 445–468. URL: https://doi.org/10.1007/s42001-020-00087-4.

[203] Sarah Masud et al. "Hate is the New Infodemic: A Topic-Aware Modeling of Hate Speech Diffusion on Twitter". In: *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. 2021, pp. 504–515.

[204] Stijn Peeters and Sal Hagen. "4CAT: Capture and Analysis Toolkit". In: *Computer Software. Vers* 1.0 (2018).

[205] Emmi Bevensee et al. "SMAT : The Social Media Analysis Toolkit". In: *Proceedings of the International AAAI Conference on Web and Social Media Workshops* (2020).

[206] Maura Conway. "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research". In: *Studies in Conflict & Terrorism* 40.1 (Jan. 2017), pp. 77–98. URL: https://doi.org/10.1080/1057610X.2016.1157408.

[207] Anh V Vu et al. "ExtremeBB: Enabling Large-Scale Research into Extremism, the Manosphere and Their Correlation by Online Forum Data". In: *arXiv preprint arXiv:2111.04479* (2021).

[208] Zorica Trajkova and Silvana Neshkovska. "Online Hate Propaganda During Election Period: The Case of Macedonia". In: *Lodz Papers in Pragmatics* 14.2 (2018), pp. 309–334. URL: https://doi.org/10.1515/lpp-2018-0015.

[209] Gobinda G. Chowdhury. "Natural Language Processing". In: *Annual Review of Information Science and Technology* 37.1 (2003), pp. 51–89. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440370103. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103.

[210] Neil J Smelser and Paul B Baltes. *International Encyclopedia of the Social & Behavioral Sciences*. eng. 1st ed. Amsterdam ; New York: Elsevier, 2001.

[211] James Powell and Matthew Hopkins. *A Librarian's Guide to Graphs, Data and the Semantic Web*. eng. Amsterdam, Netherlands: Chandos Publishing, 2015.

[212] Nasir Naveed et al. "Bad News Travel Fast: A Content-Based Analysis of Interestingness on Twitter". In: *Proceedings of the 3rd International Web Science Conference*. WebSci '11. Koblenz, Germany: Association for Computing Machinery, 2011. URL: https://doi.org/10.1145/2527031.2527052.

[213] Kateryna Lytvyniuk, Rajesh Sharma, and Anna Jurek-Loughrey. "Predicting Information Diffusion in Online Social Platforms: A Twitter Case Study". In: *Complex Networks and Their Applications VII*. Ed. by Luca Maria Aiello et al. Cham: Springer International Publishing, 2019, pp. 405–417.

[214] CLIVE SEALE and DAVID SILVERMAN. "Ensuring Rigour in Qualitative Research". In: *European Journal of Public Health* 7.4 (Dec. 1997), pp. 379–384. URL: https://doi.org/10.1093/eurpub/7.4.379.

[215] D Silverman. "Qualitative Research: Meanings or Practices?" In: *Information Systems Journal* 8.1 (Jan. 1998), pp. 3–20. URL: https://doi.org/10.1046/j.1365-2575.1998.00002.x.

[216] DataReportal. *Global Social Media Stats*. URL: https://datareportal.com/social-media-users.

[217] John Suler. "The Online Disinhibition Effect". In: *CyberPsychology & Behavior* 7.3 (2004). PMID: 15257832, pp. 321–326. eprint: https://doi.org/10.1089/1094931041291295. URL: https://doi.org/10.1089/1094931041291295.

[218] Christopher Terry and Jeff Cain. "The Emerging Issue of Digital Empathy". eng. In: *American journal of pharmaceutical education* 80.4 (May 2016), p. 58. URL: https://pubmed.ncbi.nlm.nih.gov/27293225%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4891856/.

[219] Jayne Gackenbach. *Psychology and the Internet: Intrapersonal, Interpersonal, and Transpersonal Implications*. English. Amsterdam; Boston, 2007. URL: http://site.ebrary.com/id/10160329.

[220] Cheng-Yu Lai and Chia-Hua Tsai. "Cyberbullying in the Social Networking Sites: An Online Disinhibition Effect Perspective". In: *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*. MISNC, SI, DS 2016. Union, NJ, USA: Association for Computing Machinery, 2016. URL: https://doi.org/10.1145/2955129.2955138.

[221] Terry Lee. "The Global Rise of "Fake News" and the Threat to Democratic Elections in the USA". In: *Public Administration and Policy* 22.1 (Jan. 2019), pp. 15–24. URL: https://doi.org/10.1108/PAP-04-2019-0008.

[222] Savvas Zannettou et al. "The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans". In: *J. Data and Information Quality* 11.3 (May 2019). URL: https://doi.org/10.1145/3309699.

[223] Michela Del Vicario et al. "Echo Chambers: Emotional Contagion and Group Polarization on Facebook". In: *Scientific Reports* 6.1 (2016), p. 37825. URL: https://doi.org/10.1038/srep37825.

[224] Bing He et al. “Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis”. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’21. Virtual Event, Netherlands: Association for Computing Machinery, 2021, pp. 90–94. URL: https://doi.org/10.1145/3487351.3488324.

[225] Matteo Cinelli et al. “Dynamics of Online Hate and Misinformation”. In: *Scientific Reports* 11.1 (2021), p. 22083. URL: https://doi.org/10.1038/s41598-021-01487-w.

[226] Justin Cheng et al. “Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions”. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’17. Portland, Oregon, USA: Association for Computing Machinery, 2017, pp. 1217–1230. URL: https://doi.org/10.1145/2998181.2998213.

[227] Anne Aly and Jason-Leigh Striegher. “Examining the Role of Religion in Radicalization to Violent Islamist Extremism”. In: *Studies in Conflict & Terrorism* 35.12 (Dec. 2012), pp. 849–862. URL: https://doi.org/10.1080/1057610X.2012.720243.

[228] Joseph B Walther. “Social Media and Online Hate”. In: *Current Opinion in Psychology* 45 (2022), p. 101298. URL: https://www.sciencedirect.com/science/article/pii/S2352250X21002505.

[229] Loo Seng Neo et al. “Understanding the Psychology of Persuasive Violent Extremist Online Platforms”. In: *Combating Violent Extremism and Radicalisation in the Digital Era*. IGI Global, Jan. 2016. Chap. Chapter 1, pp. 1–15.

[230] Orla Lynch. “British Muslim Youth: Radicalisation, Terrorism and the Construction of the “Other””. In: *Critical Studies on Terrorism* 6.2 (Aug. 2013), pp. 241–261. URL: https://doi.org/10.1080/17539153.2013.788863.

[231] Mudit Chaudhary, Chandni Saxena, and Helen Meng. “Countering Online Hate Speech: An NLP Perspective”. In: *arXiv preprint arXiv:2109.02941* (2021).

[232] Elissa Lee and Laura Leets. “Persuasive Storytelling by Hate Groups Online: Examining Its Effects on Adolescents”. In: *American Behavioral Scientist* 45.6 (Feb. 2002), pp. 927–957. URL: https://doi.org/10.1177/0002764202045006003.

[233] K Hazel Kwon and Anatoliy Gruzd. “Is Offensive Commenting Contagious Online? Examining Public vs Interpersonal Swearing in Response to Donald Trump’s YouTube Campaign Videos”. In: *Internet Research* 27.4 (Jan. 2017), pp. 991–1010. URL: https://doi.org/10.1108/IntR-02-2017-0072.

[234] Susan Herring et al. “Searching for Safety Online: Managing "Trolling" in a Feminist Forum”. In: *The Information Society* 18.5 (Oct. 2002), pp. 371–384. URL: https://doi.org/10.1080/01972240290108186.

[235] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. “Exposure to Hate Speech Increases Prejudice Through Desensitization”. In: *Aggressive Behavior* 44.2 (Mar. 2018), pp. 136–146. URL: https://doi.org/10.1002/ab.21737.

[236] Michael A Peters. "Limiting the Capacity for Hate: Hate Speech, Hate Groups and the Philosophy of Hate". In: *Educational Philosophy and Theory* (Aug. 2020), pp. 1–6. URL: https://doi.org/10.1080/00131857.2020.1802818.

[237] Chara Bakalis. *Cyberhate: An Issue of Continued Concern for the Council of Europe's Anti-Racism Commission*. Council of Europe, 2015.

[238] Department for Digital Culture Media and Sport (UK). *Online Harms White Paper*. United Kingdom, 2019. URL: https://www.gov.uk/government/consultations/online-harms-white-paper.

[239] Robert Gorwa, Reuben Binns, and Christian Katzenbach. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance". In: *Big Data & Society* 7.1 (Jan. 2020), p. 2053951719897945. URL: https://doi.org/10.1177/2053951719897945.

[240] Ronan Lee. "Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis". In: *International Journal of Communication* 13.0 (2019). URL: https://ijoc.org/index.php/ijoc/article/view/10123.

[241] GIFCT. *About the Global Internet Forum to Counter Terrorism*. 2019. URL: https://perma.cc/44V5-554U (visited on 02/01/2022).

[242] Zeerak Waseem et al. "Understanding Abuse: A Typology of Abusive Language Detection Subtasks". In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 78–84. URL: https://aclanthology.org/W17-3012.

[243] Mai ElSherief et al. "Peer to Peer Hate: Hate Speech Instigators and Their Targets". In: *Proceedings of the International AAAI Conference on Web and Social Media* 12 (2018), pp. 52–61.

[244] Phyllis B Gerstenfeld, Diana R Grant, and Chau-Pu Chiang. "Hate online: A Content Analysis of Extremist Internet Sites." In: *Analyses of Social Issues and Public Policy (ASAP)* 3.1 (2003), pp. 29–44.

[245] Lori Hale. "Globalization: Cultural Transmission of Racism". In: *Race, Gender & Class* 21.1/2 (Mar. 2014), pp. 112–125. URL: http://www.jstor.org/stable/43496963.

[246] Max V. Birk et al. "The Effects of Social Exclusion on Play Experience and Hostile Cognitions in Digital Games". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 3007–3019. URL: https://doi.org/10.1145/2858036.2858061.

[247] Matthew Costello and James Hawdon. "Who Are the Online Extremists Among Us? Sociodemographic Characteristics, Social Networking, and Online Experiences of Those Who Produce Online Hate Materials". In: *Violence and Gender* 5.1 (Jan. 2018), pp. 55–60. URL: https://doi.org/10.1089/vio.2017.0048.

[248] Yi-Ling Chung et al. "CONAN - COunter NArratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2819–2829. URL: https://aclanthology.org/P19-1271.

[249] Karsten Müller and Carlo Schwarz. "Fanning the Flames of Hate: Social Media and Hate Crime". In: *Journal of the European Economic Association* 19 (Oct. 2020), pp. 2131–2167. URL: https://doi.org/10.1093/jeea/jvaa045.

[250] Hitkul et al. *Capitol (Pat)riots: A Comparative Study of Twitter and Parler*. 2021. arXiv: 2101.06914 [cs.CY].

[251] Isabel Murdock, Kathleen M. Carley Carley, and Osman Yagan. "Multi-Platform Analysis of 2020 U.S. Election: Fraud and Protest Related Posts". In: *IDeaS Conference*. Pittsburgh: The Center for Informed Democracy and Social-cybersecurity), 2021.

[252] Nicholas Reimann. *Reddit Bans 'r/donaldtrump' Subreddit*. 2021. URL: https://www.forbes.com/sites/nicholasreimann/2021/01/08/reddit-bans-rdonaldtrump-subreddit/?sh=5980347038b3 (visited on 03/29/2021).

[253] WHO. *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 – 11 March 2020*. 2020. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (visited on 06/05/2022).

[254] Caroline Kantis et al. *UPDATED: Timeline of the Coronavirus*. 2023. URL: https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus (visited on 01/06/2023).

[255] Daniel Godfrey et al. *A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets*. 2014. arXiv: 1408.5427 [stat.ML].

[256] Emre Calisir and Marco Brambilla. "Wide-Spectrum Characterization of Long-Running Political Phenomena on Social Media: The Brexit Case". In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. SAC '20. Brno, Czech Republic: Association for Computing Machinery, 2020, pp. 1869–1876. URL: https://doi.org/10.1145/3341105.3374041.

[257] Keenan Jones, Jason R. C. Nurse, and Shujun Li. "Behind the Mask: A Computational Study of Anonymous' Presence on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 327–338. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7303.

[258] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *The Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[259] Jocelyn Mazarura and Alta de Waal. "A Comparison of the Performance of Latent Dirichlet Allocation and the Dirichlet Multinomial Mixture Model on Short Text". In: *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. South Africa: Institute of Electrical and Electronics Engineers (IEEE), 2016, pp. 1–6.

[260] Rishabh Mehrotra et al. "Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling". In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. Dublin, Ireland: Association for Computing Machinery, 2013, pp. 889–892. URL: https://doi.org/10.1145/2484028.2484166.

[261] James Pennebaker et al. *The Development and Psychometric Properties of LIWC2015*. Tech. rep. University of Texas, Austin, 2015.

[262] Ewa Kacewicz et al. "Pronoun Use Reflects Standings in Social Hierarchies". In: *Journal of Language and Social Psychology* 33.2 (2014), pp. 125–143. eprint: https://doi.org/10.1177/0261927X13502654. URL: https://doi.org/10.1177/0261927X13502654.

[263] James W Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us*. New York: Bloomsbury Press, 2011.

[264] Sudesh Kumar and Nancy. "Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization". In: *International Journal on Recent and Innovation Trends in Computing and Communication* 2 (2014), pp. 3161–3166.

[265] Dhaval M Dave, Drew McNichols, and Joseph J Sabia. *Political Violence, Risk Aversion, and Non-Localized Disease Spread: Evidence from the U.S. Capitol Riot*. Working Paper 28410. National Bureau of Economic Research, Feb. 2021. URL: http://www.nber.org/papers/w28410.

[266] W Phillips and R M Milner. *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online*. Polity Press, 2018.

[267] Lorraine Bowman-Grieve. "Exploring "Stormfront": A Virtual Community of the Radical Right". In: *Studies in Conflict & Terrorism* 32.11 (2009), pp. 989–1007. eprint: https://doi.org/10.1080/10576100903259951. URL: https://doi.org/10.1080/10576100903259951.

[268] Weiai (Wayne) Xu and Congcong Zhang. "Sentiment, Richness, Authority, and Relevance Model of Information Sharing During Social Crises —- The case of #MH370 Tweets". In: *Computers in Human Behavior* 89 (2018), pp. 199–206. URL: https://www.sciencedirect.com/science/article/pii/S0747563218303637.

[269] Mina Cikara, Matthew M Botvinick, and Susan T Fiske. "Us Versus Them: Social Identity Shapes Neural Responses to Intergroup Competition and Harm". In: *Psychological Science* 22.3 (Jan. 2011), pp. 306–313. URL: https://doi.org/10.1177/0956797610397667.

[270] Yael-Janette Zupnik. "A Pragmatic Analysis of the Use of Person Deixis in Political Discourse". eng. In: *Journal of Pragmatics* 21.4 (1994), pp. 339–383.

[271] Mai ElSherief et al. "Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media* 12 (2018), pp. 42–51.

[272] Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. "The Spread of COVID-19 Conspiracy Theories on Social Media and the Effect of Content Moderation". In: *Harvard Kennedy School Misinformation Review* 1.August (2020), pp. 1–19.

[273] Jacob Davey and Milo Comerford. *Between Conspiracy and Extremism: A Long COVID Threat? - Introductory Paper*. Tech. rep. Institute for Strategic Dialogue, 2021, pp. 1–14. URL: https://www.isdglobal.org/wp-content/uploads/2021/12/Between-Conspiracy-and-Extremism_A-long-COVID-threat_Introductory-Paper.pdf.

[274] Hatewatch. *Fighting AAPI Hate: Violence Against Asian Americans and Pacific Islanders Rising Amid COVID-19 Scapegoating*. 2021. URL: https://www.splcenter.org/news/2021/05/13/fighting-aapi-hate-violence-against-asian-americans-and-pacific-islanders-rising-amid-covid (visited on 07/13/2022).

[275] Lorraine Bowman-Grieve. "Exploring "Stormfront": A Virtual Community of the Radical Right". In: *Studies in Conflict & Terrorism* 32.11 (Oct. 2009), pp. 989–1007. URL: https://doi.org/10.1080/10576100903259951.

[276] Reuters. *Fact Check: Agenda 2030 is a Global Development Framework that is Not Linked to the COVID-19 Pandemic*. 2021. URL: https://www.reuters.com/article/uk-factcheck-agenda-2030-idUSKBN2AN2CQ (visited on 07/15/2022).

[277] Shahin Nazar and Toine Pieters. *Plandemic Revisited: A Product of Planned Disinformation Amplifying the COVID-19 "infodemic"*. 2021. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2021.649930.

[278] ADL. *Extremists Use Coronavirus to Advance Racist, Conspiratorial Agendas*. 2020. URL: https://www.adl.org/resources/blog/extremists-use-coronavirus-advance-racist-conspiratorial-agendas (visited on 08/15/2022).

[279] Jianshu Weng et al. "TwitterRank: Finding Topic-Sensitive Influential Twitterers". In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. WSDM '10. New York, New York, USA: Association for Computing Machinery, 2010, pp. 261–270. URL: https://doi.org/10.1145/1718487.1718520.

[280] Arlei Silva et al. "ProfileRank: Finding Relevant Content and Influential Users Based on Information Diffusion". In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. SNAKDD '13. Chicago, Illinois: Association for Computing Machinery, 2013. URL: https://doi.org/10.1145/2501025.2501033.

[281] Daniel M Romero et al. "Influence and Passivity in Social Media". In: *Lecture Notes in Computer Science*. Ed. by Dimitrios Gunopulos et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 18–33.

[282] Leon Yin. *SMAPPNYU/urlExpander: Initial release*. Aug. 2018. URL: https://doi.org/10.5281/zenodo.1345144.

[283] Weixin Wang et al. "Visualization of Large Hierarchical Data by Circle Packing". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, pp. 517–520. URL: https://doi.org/10.1145/1124772.1124851.

[284] Pasquale De Meo et al. "Generalized Louvain method for community detection in large networks". In: *2011 11th International Conference on Intelligent Systems Design and Applications*. 2011, pp. 88–93.

[285] Kate Starbird. "Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017), pp. 230–239. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14878.

[286] Savvas Zannettou et al. "Understanding Web Archiving Services and Their (Mis)Use on Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1 SE - Full Papers. AAAI, June 2018. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/15018.

[287] Tess Owen. *An Infowars #StopTheSteal Caravan Was 'Attacked' With A Milkshake*. 2020. URL: https://www.vice.com/en/article/epdxgp/an-infowars-stopthesteal-caravan-was-attacked-by-a-hero-with-a-milkshake (visited on 01/10/2021).

[288] Ben Tobin and Lucas Aulbach. *'Patriot caravan' protesting Trump's election loss heads to Washington from Louisville*. 2021. URL: https://eu.courier-journal.com/story/news/local/2021/01/05/trump-supporters-exit-louisville-for-stop-the-steal-rally/4132703001/ (visited on 01/10/2021).

[289] BBC News. *K-pop Fans Drown Out #WhiteLivesMatter Hashtag*. June 4, 2020. URL: https://www.bbc.co.uk/news/technology-52922035 (visited on 10/04/2021).

[290] Austin Hounsel et al. "Identifying Disinformation Websites Using Infrastructure Features". In: *10th USENIX Workshop on Free and Open Communications on the Internet*. USENIX. 2020.

[291] Richard Kuzma, Iain J Cruickshank, and Kathleen M Carley. "Analysis of External Content in the Vaccination Discussion on Twitter". In: *arXiv e-prints* (2021), arXiv–2107.

[292] Clyde W Holsapple, Shih-Hui Hsiao, and Ram Pakath. "Business Social Media Analytics: Characterization and Conceptual Framework". In: *Decision Support Systems* 110 (2018), pp. 32–45. URL: https://www.sciencedirect.com/science/article/pii/S0167923618300502.

[293] Douwe Kiela et al. "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2611–2624.

[294] Elizabeth Culliford and Katie Paul. *Reddit Bans 'The_Donald' Forum Amid Broad Social Media Crackdown*. 2020. URL: https://www.reuters.com/article/us-reddit-trump-idUSKBN2402K7 (visited on 02/10/2020).

[295] ADL. *ADL Debunk: Disinformation and the BLM Protests*. 2020. URL: https://www.adl.org/resources/report/adl-debunk-disinformation-and-blm-protests (visited on 02/10/2022).

[296] Alexandra Sternlicht. *Over 4,400 Arrests, 62,000 National Guard Troops Deployed: George Floyd Protests by the Numbers*. 2020. URL: https://www.forbes.com/sites/alexandrasternlicht/2020/06/02/over-4400-arrests-62000-national-guard-troops-deployed-george-floyd-protests-by-the-numbers/ (visited on 02/15/2023).

[297] Tim Chan. *K-Pop Power: Fandoms Unite to Take Over #WhiteLivesMatter Hashtag on Twitter*. 2020. URL: https://www.rollingstone.com/music/music-news/white-lives-matter-k-pop-1009581/ (visited on 02/15/2023).