Master's Degree in
Data Science and Management

Data Privacy and Security

Navigating the Digital Labyrinth: Strategies for Institutional Communication in the Era of Bot-Driven Disinformation

Supervisor
Paolo Spagnoletti

Co-Supervisor
Giuseppe Italiano

Candidate
Hanna Carucci Viterbi

Academic Year 2022/2023

To my parents and friends
that sustained me along this path

Abstract

In an era where social media platforms fundamentally alter communication dynamics, the dissemination of misinformation and disinformation has become rampant, affecting institutional communication severely. This thesis focuses on the 2016 U.S. Presidential election as a pivotal case study to explore these phenomena. Utilizing an extensive dataset of tweets, the research aims to create a policy artifact that guides institutions in countering social bots and disinformation. The artifact incorporates proactive 'prebunking' strategies and reactive measures supported by attribution-based methods. Analyzing real-world disinformation campaigns, this study seeks to enhance the integrity and transparency of institutional communication by offering a nuanced approach to managing disinformation risks.

Contents

1 Introduction **9**

2 Literature Review **11**

2.1 Dynamics and Impacts of Disinformation 12

2.2 Automated Data Processing 13

2.2.1 The rise of automated systems 13

2.2.2 Algorithms and Data Curation 14

2.2.3 Automated Systems in disinformation campaigns 14

2.2.4 Ethical and Policy implications 15

2.3 Network Dynamics in Social Media: Implications for Disinformation and Policy 16

2.3.1 Information Diffusion 17

2.3.2 Anatomy of Disinformation Campaigns 17

2.4 Institutional communication and disinformation management 18

2.4.1 The Multi-Dimensional Impact of Disinformation on Institutional Security: A Review of Recent Cases 19

2.4.2 Mitigation strategies 21

2.5 Research gaps and research question 21

3 Methodology **25**

3.1 Social Bots and Voter Influence in the 2016 U.S. Election 27

3.1.1 Data Exploration 28

4 Data Analysis **31**

4.1 Data preparation and preprocessing 32

4.1.1 The Dataset 32

4.1.2 Data Cleaning 34

4.2 Content Analysis 35

4.2.1 Language Distribution 36
4.2.2 Topics' extraction 37
4.2.3 Comparative Data Framework: An Integration of Multiple Twitter Datasets for Robust Disinformation Detection 43
4.3 Temporal Analysis 46
4.4 Network Analysis 48
4.4.1 User Metrics 48
4.4.2 Network Dynamics and Dissemination Strategies 50
4.4.3 Stratified Architectures of Information and Disinformation 51
4.4.4 Information Cascades and Amplification in Russian Twitter Networks 54
4.4.5 Network Dynamics and User Roles in Retweet Behavior 58
4.4.6 Inter-bot Interaction Analysis 60
4.5 Key Insights and Relevance for Policy Development 64

5 Discussion and Conclusions 67
5.1 Findings 67
5.2 Establishing Solution Goals and Objectives 70
5.2.1 Structure of the Policy 71
5.2.2 Structure and Distribution of Responsibilities 72
5.3 Prebunking Strategy 73
5.3.1 Amplification Metrics 75
5.3.2 Linguistic Inconsistencies 79
5.4 Monitoring and Detection Guidelines 80
5.5 Post-Incident Communication Plan 81
5.6 Post-Incident Evaluation 82
5.7 Regular Updates and Revisions 82
5.8 Policy Enforcement and Compliance 83
5.9 Inter-Departmental Synergy in Combatting Disinformation: Roles and Coordination 84
5.10 Conclusions 85
5.11 Limitations and Scope for Future Research 87

List of Figures

4.1 Language Distribution 36
4.2 LDA on Knime 37
4.3 Proportion of highest topics per account category 41
4.4 Distribution of Tweets by account category 43
4.5 Distribution of retweets by account category 43
4.6 Trolls vs. Non-Trolls features 45
4.7 Trolls vs. Non-Trolls features 46
4.8 Tweets count by date 46
4.9 Distribution of account creation years 49
4.10 Distribution of account creation months in 2014 49
4.11 Distribution of account creation months (2013-2016) for Unpopular accounts 50
4.12 Tweeting pattern over a 24 hour period for the top 20 tweeters 52
4.13 24 hour period for the top 20 most retweeted users 53
4.14 Linkage between users 55
4.15 Linkage between users 55
4.16 Linkage between users 55
4.17 Linkage between users 56
4.18 Chain Retweets 57
4.19 Chain Retweets 57
4.20 Chain Retweets 58
4.21 Directed Graph of 10 most retweeted accounts 59
4.22 Degree Distribution 62
4.23 METEVIDENCE 63
4.24 Top 10 Betweeness centrality 63

5.1 Inter-Departmental Coordination 85

List of Tables

4.1 Explanation of Features 33
4.2 Topics and Associated Words 38
4.3 Tweet Counts and Percent Changes 47

5.1 Division of Responsibilities for Managing Social Bots and Disinformation . . 74

Chapter 1

Introduction

The convergence of technology, politics, and society in the digital era has birthed a complex tapestry of phenomena that warrant nuanced scholarly investigation. Situated at this intersection, this thesis grapples with the intricate challenges presented by the proliferation of disinformation, particularly on social media platforms. While the problem has been studied in various lights—be it the role of automated data processing in manipulating narratives or the network dynamics that dictate the ebb and flow of information—this research endeavors to unify these dimensions into a holistic understanding.

The spread of disinformation has far-reaching implications that extend beyond mere misinformation; it poses risks that can have large-scale, material consequences on institutions and public safety. Take, for example, the impacts of disinformation campaigns on several significant events: the SolarWinds cyberattack and recent vulnerabilities exploited in VMWare and the subsequent denial on the TIM's outage. In each of these cases, disinformation didn't just spread falsehoods; it amplified existing crises, hindered remedial actions, and eroded public trust. By complicating these already fraught situations, disinformation escalated them into multi-dimensional crises with broader societal impacts. False narratives and misleading information hindered crisis management efforts and corroded public trust, turning a challenging technical issue into a multi-dimensional crisis.

Given this evolving landscape, with real-world, tangible ramifications, this research aims to delve deeply into the mechanics of how disinformation spreads, its impact on institutional communications, and the resultant effects on public safety and trust. By examining these factors in conjunction with one another, the study seeks to offer a more holistic understanding of the disinformation ecosystem, thereby contributing to the body of knowledge essential for developing effective countermeasures. While the main focus of the research is on analyz-

ing data from the 2016 U.S. Presidential Elections, it's important to note that the issue of disinformation is constantly evolving and expanding. The goal is to gather these insights and transform them into policy guidelines that organizations can use to combat disinformation and its harmful effects. This includes approaches that aim to educate and raise awareness among the public about the dangers of disinformation as well as reactive strategies that prioritize maintaining institutional credibility by being transparent and relying on evidence based communication in response, to incidents.

In order to proactively address disinformation, prebunking strategies are implemented with the goal of debunking false information through educational and awareness campaigns. Organizations can offer training programs, run awareness campaigns, and conduct regular briefings to empower their members with the skills needed to recognize and critically evaluate misinformation. By doing so, these initiatives help cultivate an environment that is less vulnerable to manipulation.

On the contrary, strategies for post-incident communication prioritize a prompt and transparent response to identified instances of disinformation. The implementation of timely communication, bolstered by attribution-based methods that provide substantial evidence, can effectively alleviate the influence of disinformation on institutional credibility and public trust.

Rather than isolating individual components, this thesis adopts an integrated approach. It aims to examine how algorithms and machine learning technologies, often opaque in their functioning, not only disseminate but also selectively curate information, thereby acting as potential conduits for disinformation. This is examined in tandem with how social networks, given their intricate structure and the dynamics of their information flow, can serve as both mitigators and accelerators of disinformation. These technological and social frameworks do not operate in a vacuum but are deeply embedded in real-world events and political contexts, making the impact of their confluence far more consequential. As a result, this research presents a proposal and the policy artifact remains open to ongoing enhancements and adaptations in light of emerging insights and advancements in technology.

The main objective is to provide organizations with efficient tactics for dealing with social bots and disinformation in order to ensure the authenticity and openness of their communication endeavors.

Chapter 2

Literature Review

The digital era has brought about phenomena that intertwine technology, politics and society. To fully understand the breadth and impact of these phenomena it is necessary to conduct investigations. This chapter explores the existing literature on four aspects: disinformation, automated data processing, the network dynamics in social medias and the disinformation's impact on institutional communication. Each dimension presents its complexities, which will be examined in this review to situate this study within the academic discourse.

Chapter Structure

- **Disinformation**: This section will scrutinize the dynamics at play in disinformation campaigns, particularly examining how information is manipulated and propagated. While the role of bots and algorithms is noteworthy, the focus here will shift towards understanding the efficacy of various mitigation strategies, both proactive and reactive, in curtailing the societal impact of disinformation.

- **Automated Data Processing**: This section delves into how algorithms and machine learning technologies contribute to disseminating, curating and amplifying information. It also explores the opacity of these algorithms and their sociopolitical implications.

- **Network dynamics in Social Media**: This section will delve into the intricacies of social networks, focusing on their role as platforms for information dissemination and public discourse. It will explore the structural aspects, such as nodes and connections, as well as the dynamics of information flow within these networks. The review will discuss how social networks can both mitigate and exacerbate the spread of disinfor-

mation, influencing collective behavior and decision-making in both benign and malign ways.

- **Disinformation's impact on Institutional Communication: Real-world cases and Implications**: This section delves into how organizations and public entities are particularly vulnerable to disinformation campaigns, which can severely impact their credibility and efficacy. Cases such as the SolarWinds cyberattack, false NOTAMs by the FAA and recent VMWare cyberattacks are briefly examined to illustrate the real-world implications of disinformation on institutional communication.

By exploring each theme this literature review aims to not only provide a comprehensive understanding of individual topics but also present an integrated perspective that emphasizes their connections. Moreover, this chapter sets the stage for the empirical analysis of a dataset comprising 3 million tweets from the 2016 U.S. Presidential elections, allowing for an enriched discussion of the study's findings in subsequent chapters.

2.1 Dynamics and Impacts of Disinformation

The dynamics of disinformation are intricately woven into the fabric of contemporary digital culture, greatly enabled by the amplification capacities of social media and algorithmic curation. Research indicates that disinformation campaigns often exploit psychological biases, such as confirmation bias and cognitive dissonance, to appeal to target audiences (Pennycook and Rand, 2019). Such campaigns are often multifaceted, incorporating multimedia elements like doctored images, manipulated videos, and misleading headlines, which exacerbate their virality and impact (Wardle and Derakhshan, 2017).

The impacts of disinformation are manifold and extend beyond the realm of individual cognition to influence societal and geopolitical landscapes. At the societal level, disinformation has been found to erode trust in institutions, degrade public discourse, and fuel polarization (Lewandowsky et al., 2017). It can also have tangible outcomes, influencing election results and public policy decisions, as demonstrated by the interference in the 2016 U.S. presidential election (Jamieson, 2018). Furthermore, disinformation poses unique challenges to organizations and governments, necessitating the development of new strategies and technologies for detection, mitigation, and public education. Despite ongoing efforts, there remains a significant gap in understanding how to counteract disinformation effectively in

diverse cultural contexts and among various demographic groups (Lewandowsky et al., 2017).

In conclusion, disinformation represents a complex challenge requiring interdisciplinary approaches that combine insights from psychology, data science, political science, and communication studies. While advances have been made in identifying and understanding the mechanics and impacts of disinformation, much remains to be done in developing effective counterstrategies and understanding their long-term efficacy.

2.2 Automated Data Processing

2.2.1 The rise of automated systems

The advent of big data and computational power has marked a seismic shift in the landscape of information management and analytics. Automated data processing, encompassing a broad array of techniques from data mining (Han et al., 2011), machine learning (Goodfellow et al., 2016), to natural language processing (Jurafsky and Martin, 2019), has become critically central to a variety of applications including, but not limited to, social media platforms, online information dissemination, and even real-time decision-making systems (Kitchin, 2014; Mayer-Schönberger and Cukier, 2013).

This surge in automated data processing has profound implications for information asymmetry, effectively altering how information is curated, distributed, and consumed (Boyd and Crawford, 2012). Particularly, the algorithms employed in these processes have generated new paradigms in personalized content delivery (Pariser, 2011), thereby influencing user behavior and public opinion (Tufekci, 2015).

The complexity of these algorithms often renders them as "black boxes" (Pasquale, 2015), whose inner workings are incomprehensible to end-users but whose impact on information ecosystems is substantial and far-reaching (Gillespie, 2014).
Moreover, the rise of automated data processing tools has raised ethical and governance challenges, emphasizing the need for responsible AI and algorithmic accountability (Zarsky, 2016; Mittelstadt et al., 2016). Questions surrounding data privacy (Nissenbaum, 2010), discrimination (Barocas and Selbst, 2016), and even national security (Schneier, 2015) are increasingly being scrutinized as these automated systems become deeply embedded in our social and institutional fabric.

The integration of automated data processing with other evolving phenomena, like social bots and disinformation campaigns, underscores its dynamic and multifaceted role in the contemporary information age (Woolley and Howard, 2016; Ferrara et al., 2016). As such, understanding this landscape requires an interdisciplinary approach, fusing insights from computer science, social sciences, and policy studies to fully grasp the intricate web of implications (O'Neil, 2016; Diakopoulos, 2015).

2.2.2 Algorithms and Data Curation

Algorithms play a crucial role in sorting, filtering, and presenting data to end-users in a manner that is ostensibly objective and relevant (Diakopoulos, 2015). These algorithms are often seen as neutral arbiters of information; however, their decision-making processes are frequently opaque, making them largely inscrutable to both end-users and regulators (Gillespie, 2014). This lack of transparency not only creates a "black box" effect but also opens the door to biases that could be encoded into these algorithms, either inadvertently or by design (Eslami et al., 2015).

Moreover, the proprietary nature of these algorithms, especially those employed by large tech companies, often precludes public scrutiny. As a result, the algorithms' real-world impacts—ranging from reinforcing existing social and cultural biases to affecting democratic processes—remain underexamined. This scenario creates an accountability vacuum, raising pressing questions about governance and ethical considerations. Therefore, while algorithms bring efficiency and scalability to data processing, their opaqueness and potential for bias pose challenges that call for rigorous interdisciplinary scrutiny.

2.2.3 Automated Systems in disinformation campaigns

The emergence and proliferation of social bots in disinformation campaigns have raised concerns among researchers and policymakers alike (Ferrara, 2020). These automated software agents can rapidly spread false narratives by generating, sharing, and interacting with content, effectively manipulating public discourse and opinion (Shao et al., 2018). The presence of social bots on popular platforms like Twitter, Facebook, and Instagram has only intensified the challenge of combatting disinformation, especially during times of crisis (Bessi and Ferrara, 2016).

Several studies have demonstrated that social bots contribute significantly to the dissemination of false information and can even outpace human users in sharing misleading content (Vosoughi et al., 2018). By exploiting the algorithms that govern social media platforms, these bots can quickly amplify their reach and influence, increasing the visibility of disinformation and making it more difficult to debunk (Gorwa and Guilbeault, 2020).

Social bots can also operate in a more targeted manner, engaging specific communities or demographics to create division and discord (DiResta et al., 2019). For instance, bots may exploit existing social or political divides to spread disinformation, further polarizing groups and undermining collective efforts to address a crisis (Badawy et al., 2018). Additionally, social bots can be used to attack the credibility of authoritative sources, further complicating the efforts of institutions to effectively communicate with the public during emergencies (Woolley and Guilbeault, 2017).

In response to the growing threat posed by social bots, researchers and technologists have begun to develop tools and strategies to identify and mitigate their impact on the spread of disinformation (Ferrara, 2020). These efforts include machine learning algorithms that can detect bot-like behavior, as well as public awareness campaigns aimed at educating users on how to recognize and report suspicious accounts (Davis et al., 2016). Despite these advances, the rapidly evolving nature of social bots and their ever-increasing sophistication present ongoing challenges for researchers, policymakers, and platforms in managing post-incident disinformation (Zannettou et al., 2019).

2.2.4 Ethical and Policy implications

The burgeoning landscape of automated data processing is not without its ethical quandaries and policy challenges. At the heart of the issue lies the question of data privacy (Zuboff, 2019; Solove, 2008). As algorithms become more sophisticated in their data-mining capabilities, concerns over unauthorized data access and misuse escalate (Nissenbaum, 2010). The matter is further complicated by the commodification of user data, which is often harvested and sold to third parties without the explicit consent of users (Couldry and Mejias, 2019).

Moreover, the automated systems and algorithms that are designed to sift through and

disseminate information have the potential to be exploited for spreading disinformation and influencing public opinion (**?**). Their ability to amplify certain narratives over others raises questions about their role in shaping the political and social discourse, making algorithmic governance a topic of urgent public interest (Gillespie, 2018).

Finally, these ethical challenges spill over into the realm of policy-making, requiring governments and international bodies to grapple with issues like antitrust regulations, data protection laws, and the ethical considerations of machine autonomy (Cath et al., 2018; eu2, 2018). As these automated systems become increasingly integral to both individual lives and societal structures, it is imperative to address these ethical and policy implications in a comprehensive, interdisciplinary manner (Mittelstadt et al., 2016).

2.3 Network Dynamics in Social Media: Implications for Disinformation and Policy

Social media platforms operate on complex networks that interconnect millions of users worldwide. Understanding the topology of these networks is crucial to gaining insights into how information, and consequently disinformation, spreads (Tucker et al., 2018). Key metrics such as degree distribution and clustering coefficients offer an analytical framework for investigating the dynamics of these networks (Starbird, 2019). One major concern arises from the algorithms that prioritize certain content over others. These algorithms not only influence public opinion but also raise serious issues related to data privacy and cybersecurity (Tucker et al., 2018). The significance of measuring centrality in these networks cannot be overstated; influential nodes often act as major sources or amplifiers of information, including disinformation (Starbird, 2019). The phenomena of echo chambers and filter bubbles further complicate the landscape, leading to increased political polarization (Tucker et al., 2018). Moreover, the network-based tactics used for spreading disinformation make it necessary for a multidisciplinary approach to policy-making, incorporating fields ranging from IT to social psychology. In the context of the U.S. presidential elections, these network phenomena have shown a demonstrable impact on electoral outcomes and have been leveraged for both spreading and combating disinformation (Starbird, 2019; Tucker et al., 2018).

2.3.1 Information Diffusion

Understanding how information, including disinformation, circulates within social media networks is a central focus of current academic inquiry. Research in this area is often multidimensional, exploring both algorithmic and human factors that affect the speed and scope of information spread. One significant aspect under study is the role of platform-specific algorithms in amplifying or suppressing certain types of content. These algorithms determine what appears in a user's feed, taking into account various factors such as prior user engagement, the popularity of the post, and its perceived relevance to the user (Bakshy et al., 2012).

Simultaneously, human factors play an indispensable role in this ecosystem. User behavior, such as 'liking,' 'sharing,' or commenting, serves as a form of social signaling and impacts the algorithmic dissemination of posts (Bakshy et al., 2012). The synergistic relationship between user actions and algorithmic functions creates a complex environment for the spread of information, and, crucially, disinformation. Consequently, understanding this relationship is not only important for academics but also for policymakers aiming to craft effective strategies for mitigating the negative impacts of disinformation.

Social media platforms often function as echo chambers, where users are exposed primarily to opinions and facts that align with their existing beliefs (Pariser, 2011). This phenomenon exacerbates the spread of disinformation as users are more likely to share information that aligns with their preconceived notions, whether or not that information is accurate. The literature suggests that these echo chambers could be contributing to societal polarization, as individuals become more entrenched in their views when only exposed to confirming information (Pariser, 2011).

Despite these insights, gaps remain in the existing research. Most notably, the literature often lacks empirical evidence about the effectiveness of potential interventions aimed at halting the spread of disinformation. Further studies are needed to address these limitations and to offer more concrete guidance for policymakers.

2.3.2 Anatomy of Disinformation Campaigns

Understanding the anatomy of disinformation campaigns has become a focal point for contemporary research, which includes various elements such as the origin, mechanisms of dissemination, and the ultimate impact on public discourse. Tucker et al. (2018) provided a comprehensive framework for how social media platforms could be exploited for disinforma-

tion, emphasizing the role of amplification metrics like hashtags and mentions in the spreading of misleading narratives. Alongside this, Vosoughi et al. (2018) reported that falsehoods are 70% more likely to be retweeted than the truth, a statistic that magnifies the potential for disinformation to spread widely.

Researchers like Starbird (2019) underscore the role of bot accounts in disseminating false narratives, while Marwick and Lewis (2017) delve into how certain communities become echo chambers, thus serving as fertile grounds for the spread of disinformation. Botometic analyses have even identified temporal patterns of disinformation, linking spikes in misleading narratives to real-world events or trending topics Ferrara (2020).

Another intriguing line of inquiry is the study of network structures. According to Krebs (2002), disinformation often exploits 'structural holes' in social networks to facilitate its spread. Moreover, Allcott and Gentzkow (2017) have looked into how the spread of fake news can be associated with polarization, thus perpetuating a cycle where echo chambers and algorithmic filtering further divide the public discourse.

The nuanced understanding of these various components is essential for policymakers, technology companies, and civil society organizations in creating effective strategies for counteracting disinformation. For example, leveraging algorithmic solutions to identify 'superspreaders' of false narratives (Lazer et al., 2018) or implementing digital literacy programs that equip users with the skills to recognize misleading content (Lewandowsky et al., 2017).

Given the multifaceted challenges that disinformation presents, it's crucial that ongoing research continues to dissect its anatomy to offer data-driven strategies for mitigation. Such an understanding serves as a foundation for creating policies that are not just reactionary but preemptive in neutralizing disinformation campaigns.

2.4 Institutional communication and disinformation management

During crises, it is crucial for institutions to engage in efficient communication practices as a means of managing disinformation. According to Lachlan et al. (2016), the provision of clear, consistent and timely messaging from authorities serves to uphold public trust and combat

false narratives. Similarly, Reynolds and Seeger (2005) have emphasized the significance of such messaging in maintaining credibility during critical situations. Furthermore, recent research has underscored the necessity for institutions to conduct social media monitoring while also actively engaging with users through rapid response mechanisms that counteract disinformation efforts (Graham and Avery, 2013).

Institutions can attain efficacious crisis communication by prioritizing the cultivation of transparency and credibility in their messaging, which could diminish public uncertainty and doubt (Lachlan et al., 2016). It is advantageous for institutions to interact with the public proactively through social media platforms to recognize and manage developing concerns or rumors before they reach a critical stage (Graham and Avery, 2013). Moreover, partnering with established media channels and influential persons may aid in promoting accurate information while refuting misinformation that has potential consequences on prevailing attitudes or behaviors among the masses (Veil et al., 2011).

In addition to addressing disinformation, effective institutional communication also involves providing actionable guidance and resources for the public to navigate the crisis situation (Reynolds and Seeger, 2005). By demonstrating empathy, compassion, and understanding of the public's concerns, institutions can strengthen their relationship with the public and promote a sense of collective resilience and trust.

Ultimately, effective institutional communication during crises requires an adaptive and dynamic approach that takes into account the evolving nature of the crisis and the information landscape. By employing clear, consistent messaging, engaging with the public, monitoring social media, and responding rapidly to disinformation, institutions can minimize the negative consequences of misinformation and foster a more informed and resilient public response.

2.4.1 The Multi-Dimensional Impact of Disinformation on Institutional Security: A Review of Recent Cases

Examples of how automated accounts can compound existing vulnerabilities within institutions include the SolarWinds breach and the more recent TIM outage in Italy. The advent of automated disinformation operations made managing these crises more difficult even though they were largely technological in origin and featured security breaches and service interruptions. These automated accounts show up as key catalysts that complicate institutional

responses rather than just being minor participants.

In these instances, an additional layer of uncertainty and misinformation exacerbated the primary problem - a cybersecurity attack. This not only obstructs immediate corrective actions but also calls into question the institutions' credibility, which is essential to their long-term effectiveness and public trust. Therefore, automated disinformation efforts present organizations with a twin quandary: the urgent requirement to manage public image and information flow as well as the immediate need to address any technical or security issues.

Even if these occurrences might not be considered "disasters" in the conventional sense, they certainly highlight the difficulties that governments and organizations must deal with in the digital age. They emphasize the growing significance of incorporating public trust management and communication tactics into our crisis response frameworks.

For instance, the SolarWinds cyberattack in December 2020 significantly compromised key U.S. government departments and Fortune 500 companies. The situation was further complicated by the deliberate spread of misinformation aimed at confusing the attribution of the attack and minimizing its impact, thereby impeding remedial actions (Committee, 2021).

Similarly, there was an hacker attack in Italy on the on February 2023 that caused an outage of TIM services. The attack was reportedly carried out by a group of hackers who demanded a ransom payment in exchange for restoring the services. The attack affected millions of TIM users and caused significant disruption to the company's operations. The hashtag timdown has been used on social media to discuss the outage and share information about the situation. (Tod, 2023) The cyberattacks on VMWare presented a two-fold crisis. On one level, there was the immediate threat from the exploitation of system vulnerabilities. On another level, automated accounts propelled disinformation campaigns that created confusion around the remedial actions to be taken. These disinformation campaigns effectively served to paralyze or slow down response efforts, making it difficult for both institutions and the public to discern appropriate action steps (Reuters, 2023)

The role of automated accounts here is not just peripheral; they are central actors in the theater of modern crises, capable of generating significant consequences. Thus, they present a new kind of challenge that institutions must grapple with: the need to manage not just the immediate technical or security issues at hand but also a parallel crisis in public communi-

cation and trust, often magnified and mutated by automated disinformation campaigns.

2.4.2 Mitigation strategies

Prebunking is a proactive approach, where the aim is to inoculate the public against disinformation by exposing them to a diluted form of the misleading argument, essentially 'vaccinating' them against the falsehood before they encounter it. This technique relies on cognitive psychology and aims to build resistance in the audience's mind. Studies such as van der Linden et al. (2020) demonstrate its efficacy, particularly when implemented via interactive online platforms.

On the other hand, post-incident techniques include fact-checking, digital forensics, and public corrections. These reactive measures, which come into play after the disinformation has already spread, have their own set of challenges and advantages. According to Lewandowsky et al. (2017), post-incident corrections can be effective but are often hampered by the 'continued influence effect,' where discredited information continues to influence opinions. Additionally, these methods may suffer from the 'backfire effect,' where the act of correcting misinformation ironically ends up reinforcing the false belief among a subset of the audience.

However, institutions are becoming increasingly savvy in employing a mix of both strategies. For instance, real-time fact-checking, an innovative fusion of prebunking and post-incident techniques, provides immediate corrections during live events, thereby combining the benefits of both approaches. Despite the advancement in these mitigation techniques, the literature reveals gaps in understanding the long-term efficacy of these approaches and how they interact with variables like demographic factors and pre-existing beliefs.

Thus, while both prebunking and post-incident techniques have proven to be useful tools in the fight against disinformation, their efficacy varies and is influenced by several external factors, including the speed at which the false information is spreading and the audience's pre-existing beliefs and biases. Further research is needed to optimize these strategies for diverse information environments and demographic groups.

2.5 Research gaps and research question

The existing body of research has made significant strides in understanding the role of automated systems in the spread of disinformation (Ferrara et al., 2016; Shao et al., 2018).

However, the field remains less explored in terms of identifying the tactical, temporal, and network patterns that contribute to the evolution and rapid dissemination of false narratives. While Starbird et al. (2018) and Marwick and Lewis (2017) have examined the use of social media in information campaigns, they have not extensively delved into the specific tactics and strategies employed in the process. Temporal patterns, or the fluctuation of disinformation campaigns in response to real-world events, have been noted but not thoroughly examined in their contribution to the efficacy of such campaigns (Vosoughi et al., 2018).

Moreover, the literature often overlooks the nuances in network behavior that could provide crucial insights into countering disinformation effectively (Conway et al., 2017). This is particularly problematic given that the morphing landscape of social media platforms and the actors involved require more nuanced and adaptable policy frameworks (Edwards et al., 2018). In terms of ethical and policy implications, there's substantive discourse (Zuboff, 2019; O'Neil, 2016), yet a comprehensive, data-driven, and adaptable policy framework specifically tailored to counter disinformation is not well articulated in existing studies (Bodo et al., 2019).

This gap becomes even more crucial as policymakers and institutions confront the ever-evolving challenges of identifying, understanding, and countering disinformation campaigns. The urgency to address this gap is not just academic but of high social relevance, especially when considering the rising global implications of disinformation on democratic processes (Woolley and Howard, 2016; Tucker et al., 2018).The aim of the study is to bridge these gaps by providing actionable insights into the tactics, temporal patterns, and network behaviors that shape the spread of disinformation, offering a foundation upon which to build proactive and reactive policy measures. While the literature has extensively examined the role of social bots in disinformation campaigns and the importance of effective institutional communication in addressing disinformation, there is limited research on the specific strategies institutions can adopt to counteract the influence of social bots in post-incident communication. Additionally, little is known about the potential collaboration between institutions and social media platforms in detecting and mitigating the impact of social bots during crises, as well as the role of public engagement in combating disinformation driven by social bots. Based on these gaps in the literature, the following research questions are proposed:

- *How can institutions effectively manage post-incident disinformation by mitigating the influence of social bots in their communication strategies?*
- *How can public engagement be incorporated into institutional communication strategies to combat disinformation driven by social bots?*

- *What strategies and mechanisms contribute to the spread of disinformation in digital environments, and how can empirical data inform the development of more effective policy interventions?*

Addressing these research questions will provide valuable insights into the development of communication strategies that leverage the positive potential of social bots, address the challenges of disinformation management in post-incident situations, and explore collaborative efforts between institutions, social media platforms, and the public. This research will contribute to a more comprehensive understanding of the dynamic interplay between various stakeholders in crisis communication, ultimately leading to more effective approaches in managing disinformation and promoting accurate information dissemination.

Chapter 3

Methodology

In this chapter, an in-depth analysis of the methodology utilized in conducting this research study is presented. The main objective of this investigation is to address and resolve the issue of social bots' influence on institutional communication following crisis incidents. In this particular context, crisis incidents refer to various situations where institutions encounter substantial obstacles that demand an urgent or high-level reaction. These scenarios can involve natural calamities, cyber assaults, public health crises, financial misdeeds, or prominent disputes. The specific characteristics of the crisis would determine the appropriate mode of communication needed for effective response and recovery. Moreover, it would also influence the target audience as well as potentially shape the involvement of social bots in managing and mitigating its impact. There are several reasons why studying crisis incidents is important. Firstly, these occurrences are typically characterized by a high level of uncertainty and rapidly changing information. Therefore, it is crucial for institutions to communicate effectively in order to manage the situation and uphold public trust. Secondly, crises often garner significant attention from both the public and media, which can result in increased scrutiny and pressure for institutions to respond appropriately. In the modern era of technology, social bots have emerged as important players in shaping the information landscape on social media platforms. These automated entities are capable of generating content and engaging with human users. During times of crisis, their presence becomes especially significant. Social bots have the power to manipulate the dissemination of both accurate information and misinformation, which in turn affects how the general public perceives a crisis situation, evaluates an institution's response to it, and ultimately influences the overall outcome.
In the context of crisis situations, institutions often face challenges in effectively managing the rapid dissemination of information, which can sometimes include false or misleading

content. Gaining a comprehensive understanding of how social bots impact this process is crucial for developing more efficient communication strategies. This entails exploring methods to accurately anticipate and mitigate the actions carried out by malicious social bots, as well as potentially leveraging these automated tools positively to distribute reliable and accurate information that counters misinformation. The study's emphasis on prebunking, which involves actively debunking potential misinformation before it becomes widespread, and attribution-based responses, which involve identifying and exposing the sources of misinformation, indicates its pertinence in the context of disinformation campaigns during crises. In such situations, social bots may be employed to deliberately disseminate false or misleading information. Gaining insights into these dynamics can prove pivotal for institutions aiming to foster trustworthiness, effectively handle public perceptions, and achieve a successful resolution amidst a crisis.

This research paper is intentionally organized as an epxlorative study with the goal of understanding and clarifying this complex and rapidly evolving phenomenon. At the core of this work is an exploration of a dataset that encompasses dimensions related to the phenomenon being studied. This phenomenon exists at the intersection of automated data processing, the dynamics of networks and the wide ranging impacts of disinformation. Each of these areas offers a rich vein of insight and is targeted for deep investigation in the revised literature review. The research plan is structured to unfold across three distinct but interrelated analytical areas: content analysis, temporal analysis, and network analysis:

- **Content Analysis**: This area aims to decode the textual and contextual elements in the dataset. By examining how disinformation is framed and conveyed, the aim is to understand the psychological and cognitive mechanisms that make such messages persuasive or influential.

- **Temporal Analysis**: This examines the timing and frequency of disinformation campaigns. The objective is to identify patterns or cycles, such as whether disinformation spikes during politically sensitive periods or crises, thereby offering insights into its strategic deployment.

- **Network Analysis**: This area investigates how disinformation spreads across social networks. By examining the topology and flow of misleading information, it's possible to identify key nodes or influencers who might serve as accelerants for disinformation.

By adopting this tripartite analytical framework, the research aims to offer a more comprehensive view of the disinformation landscape. Furthermore, the integration of these analytical

areas enables a holistic understanding of how disinformation operates, evolves, and impacts society.

After the empirical data is analyzed and interpreted in light of these three analytical areas, the study will synthesize these findings to align them with existing literature. The objective here is not only to validate or challenge existing theories but also to extend the academic discourse by contributing new, data-driven insights.
In the concluding sections, policy implications will be elaborated upon. These insights will serve as actionable recommendations that can help organizations and policymakers to counteract the adverse impacts of disinformation more effectively. Given how quickly disinformation and technology evolve combining findings, with established theories will both validate existing understandings and present new perspectives that push forward academic discourse.

3.1 Social Bots and Voter Influence in the 2016 U.S. Election

The role of social bots during the 2016 U.S. presidential election warrants critical examination for its influence on the political landscape. This phenomenon constitutes a significant intersection between technology and politics, where automation transcends mere computational convenience to become a tool for shaping public opinion. The existing literature broadly agrees on the bots' impact on amplifying divisive and polarizing narratives. A seminal study by Howard and Kollanyi (2016) posited that social bots favored then-candidate Donald Trump significantly more than Hillary Clinton in disseminating information. This was not merely an innocuous propagation of information but a form of algorithmic governance that guided political discourse in a particular direction, creating what some researchers term as "manufactured consensus" (Woolley and Guilbeault, 2017).

In a climate already marked by political polarization, these bots acted as catalysts for discord, targeting hot-button issues such as immigration, gun control, and healthcare. Far from being neutral conduits of information, they were strategic in their orientation, aimed at not just generating content but also at steering the conversations towards particular outcomes, effectively biasing the democratic process (Bessi and Ferrara, 2016). These bots, by generating hyper-partisan content, created echo chambers, thereby decreasing the potential for constructive, bipartisan dialogue.

3.1.1 Data Exploration

In order to gain a thorough understanding of the dynamics and implications of social bots during crisis incidents, this study's methodology begins by conducting an extensive analysis of the 2016 U.S. presidential election. This particular event is chosen as a pivotal case study due to its wealth of empirical data available for examination. During this phase, the main goal is to carefully examine the tactics employed by social bots in order to disseminate false information and manipulate public discussions, ultimately influencing how institutions communicate. In order to accomplish this objective, the investigation will concentrate on different aspects including the extent and characteristics of social bot activity, the kinds of misinformation that are being spread, which platforms are predominantly targeted for dissemination purposes, as well as determining which demographic groups are most affected. The 2016 U.S. Presidential Election has been recognized as an important case study for several compelling reasons. Firstly, it was a highly significant event that garnered substantial global attention and represented a turning point in the American political landscape. This election highlighted the increasing influence of digital platforms in shaping public opinion. Secondly, it brought to light the role played by social bots in spreading false information and impacting public discourse. Thorough analyses conducted after the election revealed widespread use of these automated entities on social media platforms with strategic efforts to manipulate public sentiment, amplify divisive issues, and cultivate an atmosphere of distrust.

In addition, the range of false information spread by social bots during this time was diverse and included everything from political propaganda to fabricated news. This creates a complex landscape that requires careful examination of the strategies employed by these entities. It is also worth mentioning that analyzing how institutions responded to misinformation driven by bots offers an important opportunity to evaluate the effectiveness of existing communication strategies and identify any gaps or shortcomings.

This in-depth exploration of the 2016 U.S. Presidential Election will provide valuable insights into how social bots operate during high-stakes events and offer a deeper understanding of the challenges associated with managing and mitigating their impact on society.

The primary objective of this research study is to gain a comprehensive understanding and address the intricate involvement of social bots in institutional communication following crises. A key focus is exploring strategies that can effectively counteract the spread of dis-

information. In today's digital world, social bots, which are automated entities programmed to mimic human behavior, are prevalent. Their presence becomes particularly pronounced during times of crisis as they are frequently utilized to propagate false narratives and misleading information.

In the face of increasing disruptive incidents, effectively dealing with misinformation in uncertain situations is becoming a major challenge. If disinformation during crises is not properly managed, it can undermine disaster response efforts, intensify public distress, and erode trust in important institutions.

In order to create communication strategies that effectively address the negative effects of disinformation on public safety and well-being, it is essential to comprehend these dynamics. With this objective in mind, the research investigates various methods for addressing misinformation during crisis situations, with a particular emphasis on the concept of 'prebunking' as a proactive approach. Prebunking involves taking action against potential misinformation before it spreads widely, which can be seen as just as important as managing communication after an incident occurs.

An in-depth examination of the impact of social bots on communication following an incident will serve as a foundation for suggesting approaches to mitigate the spread of false information during crises. Taking a comprehensive view that spans before, during, and after a crisis allows to identify both opportunities and challenges regarding collaboration between institutions, social media platforms, and the public. Each of these stakeholders plays a vital role in combatting misinformation. By thoroughly analyzing every stage of a crisis, starting from its initiation until resolution, it's possible to gain valuable insights into developing robust strategies that effectively tackle issues related to disinformation.

To summarize, although communicating after a crisis is crucial for handling and minimizing its impact, taking proactive steps such as prebunking also plays a vital role in preventing the spread of misinformation from the outset. Both preventive measures and responsive actions are indispensable components in combating disinformation during times of crises.

Chapter 4

Data Analysis

The upcoming analysis undertakes a thorough investigation of the behavior of social bots, with a specific focus on identifying distinct patterns and strategies used during critical political events. This examination holds great importance in today's digital environment, as disinformation and the manipulation of public discourse continue to be pressing issues. The study is organized to explore three key aspects of this phenomenon: an examination of temporal changes and trends over time; an analysis of content using Natural Language Processing techniques to identify linguistic nuances and thematic orientations; and an exploration of network patterns, looking closely at interaction behaviors within the social media ecosystem. An analytical approach is utilized to uncover the intricate mechanisms employed by social bots, which is crucial in developing effective strategies against them. The temporal analysis focuses on mapping the progression of bot activities over time, identifying any patterns or variations related to specific events. Content analysis goes beyond surface-level examination and delves into semantic structures, sentiment patterns, and rhetorical devices commonly associated with bot-generated content. Additionally, network analysis explores interconnectedness and clustering phenomena to shed light on coordinated behaviors.

It is important to emphasize that the underlying methodologies adhere to rigorous analytical standards, employing sophisticated algorithms, statistical models, and machine learning techniques to obtain detailed insights. Additionally, a dedicated section will address data preparation and cleansing processes which are crucial for ensuring the reliability and accuracy of the findings.

The objective of this research effort extends beyond academic pursuits; it serves as a fundamental contribution towards formulating effective policy measures. The derived insights have the potential to significantly enhance our understanding of social bots' impact on digital

communication, providing empirical evidence for guiding future interventions in this dynamic domain.

4.1 Data preparation and preprocessing

4.1.1 The Dataset

The information used for this analysis comes from a large collection of around three million tweets. These tweets were compiled and made public by the respected analytical journalism organization called FiveThirtyEight. On July 25, 2018, FiveThirtyEight acquired data from Clemson University researchers Darren Linvill and Patrick Warren. These two individuals, an associate professor of communication and an associate professor of economics respectively, collected the data by utilizing custom searches through Social Studio. This tool is owned by Salesforce and was specifically contracted for use at Clemson's Social Media Listening Center. The dataset is comprehensive as it includes various important details for every tweet such as the author's name, the actual text of the tweet, the date when it was posted, the number of followers that the author had at that time, how many accounts they were following and whether or not it was a retweet. These metadata attributes a multidimensional analysis of social bot activity by examining their behaviors, strategies and patterns of interaction.
The dataset includes tweets from February 2012 to May 2018, with the majority of tweets falling between 2015 and 2017. This specific time period is significant due to its coverage of the 2016 U.S. presidential elections, which serves as a central focus for the case study. During this crucial event, social bots played a prominent role in shaping public discourse, making it an ideal environment to explore the complexities of bot-driven communication during crises. The wealth of data available allows for a comprehensive analysis that sheds light on the strategies employed by social bots to manipulate public opinion and provides valuable insights into how institutions can effectively counter their influence. Knowledge gained from studying this timeframe will contribute towards developing robust and resilient communication strategies necessary for managing and combating the impact caused by social bots.
Therefore, this dataset becomes an invaluable resource for conducting research in this field. The data set has been carefully curated and serves as a valuable resource for investigating the role and influence of social bots in digital communication.

The files have the following columns:

Definition	**Feature Explanation**
external_author_id	An author account ID from Twitter
author	The handle sending the tweet
content	The text of the tweet
region	A region classification, as determined by Social Studio
language	The language of the tweet
publish_date	The date and time the tweet was sent
harvested_date	The date and time the tweet was collected by Social Studio
following	The number of accounts the handle was following at the time of the tweet
followers	The number of followers the handle had at the time of the tweet
updates	The number of "update actions" on the account that authored the tweet, including tweets, retweets and likes
post_type	Indicates if the tweet was a retweet or a quote-tweet
account_type	Specific account theme, as coded by Linvill and Warren
retweet	A binary indicator of whether or not the tweet is a retweet
account_category	General account theme, as coded by Linvill and Warren
new_june_2018	A binary indicator of whether the handle was newly listed in June 2018
alt_external_id	Reconstruction of author account ID from Twitter, derived from article_url variable and the first list provided to Congress
tweet_id	Unique id assigned by Twitter to each status update, derived from article_url
article_url	Link to the original tweet. Now redirects to the "Account Suspended" page
tco1_step1	First redirect for the first http(s)://t.co/ link in a tweet, if it exists
tco2_step1	First redirect for the second http(s)://t.co/ link in a tweet, if it exists
tco3_step1	First redirect for the third http(s)://t.co/ link in a tweet, if it exists

Table 4.1: Explanation of Features

4.1.2 Data Cleaning

To begin the analysis, it was necessary to combine the initially separate datasets. Initially, there were 12 different datasets that contained specific segments of data from various time periods within the chosen timeframe (February 2012 - May 2018). It was crucial to consolidate these datasets into one comprehensive dataset in order to fully comprehend the activities of social bots throughout the entire duration being studied. By combining all of this information it was possible to carry out a continuous examination of bot actions, patterns of influence. How they affected communication, within institutions during the crucial 2016 U.S. Presidential elections. As part of the data preparation stage all mentions have been extracted from the tweets to conduct an analysis of interactions involving bots. The extraction process used was based on the list of tweets obtained from the dataset. To focus on the columns for the study and streamline the data excluded several columns. These excluded columns included "Unnamed; 0" "external author id" "tco1 step1" "harvested date" "following" "updates","post type","account type","new june 2018" "alt id","tweet id","article url" "tco2 step1" "Tco3 step1". During the phase it was extensively explored the dataset to uncover insights related to social bot activity and its impact, on public discussions during the 2016 U.S. Presidential elections. This involved using methodologies to identify patterns and interactions within the data. To begin this process it was successfully identified user mentions in tweet content by looking for "@" symbols, which indicate references to users. This thorough examination provided a nuanced understanding of how users mentioned each other and the complex dynamics of communication between accounts. Additionally it was effectively examined hashtags (denoted by "#") used in tweet content shedding light on relationships and conversations, among platform users. This exploration notably illuminated the relationships and dialogues interwoven among different users within the platform.
Through this analytical lens, a holistic overview of the central subjects dominating discussions was unveiled, enabling a comprehensive grasp of the pivotal themes driving public discourse.

Furthermore, an integral facet of the analysis entailed considering the number of followers attributed to each account at the time of tweeting. This strategic step bolstered the analysis by pinpointing influential users whose messages potentially resonated with a broader audience. Gathering stats about social media followers helped sketch out a visual map, showcasing how far a message could ripple across the vast online landscape. What's more, by tracking the tick-tock of retweets, the study cracked open the vault on how info from other users zipped across the feed. This deep dive offered up a treasure trove on why certain

posts grabbed eyeballs and buzz. At the heart of this number-crunching lay the effort to spell out and put on display the intricate web that connects user mentions, popular hashtags, and followers, as well as retweet action. Contrary to being informal or non-academic, these approaches revealed the exchange of information among users, especially during critical events such as the U.S. 2016 election period. The findings threw back the curtain on the role of bots and users alike in steering the conversation and setting the digital agenda. Alongside an exploration of the impact of bots and algorithms, on discourse the study also presented an array of visually appealing charts and concrete statistical data. These were no mere window dressing—they peeled back the layers on the knotty interactions during that fateful election season. Using easy-to-grasp visuals, the study broke down recurring user habits, hashtag faves, and the scoop on who was sharing what. .

Moreover the study included a range of metrics based on numbers to provide an understanding of the influential figures, in the digital world. It wasn't solely focused on follower counts or retweet tallies. By conducting research it was possible to identify the frequently mentioned individuals and how often content was shared, basing insights on solid numerical data. This meticulous analysis of interactions helped determine the relevance and importance of users, in the online sphere.

The amalgamation of visual elucidation and analytical trends not only deepened the intricacy of the analysis but also engendered a more comprehensive comprehension of the roles enacted by automated accounts and human users alike, in shaping the digital discourse during the elections. By synergizing qualitative interpretations with quantitative evaluations, this segment of the analysis encapsulated the intricate essence of digital dialogues, thus contributing to a more holistic and nuanced interpretation of the dataset.

4.2 Content Analysis

Content analysis utilizes techniques in natural language processing to analyze the textual elements of a dataset. Through the application of different algorithms, recurring themes, keywords, and sentiments are identified. This process provides insights into the prevailing discourse and places particular emphasis on determining central topics that attract bots' attention.

4.2.1 Language Distribution

The analysis of language distribution provides insights into the research questions. English is the language in the dataset making up a 76%. This aligns with the prevalence of English on social media platforms. Italian accounts for 12% of the dataset indicating a focus on interactions in Italian. With 8% representing Russian it suggests a presence that could be related to contexts or online communities in Russian speaking regions. The remaining 3.3% attributed to languages showcases the diversity within the dataset. The prominence of English suggests that this study could have a cultural impact due to its global reach. The substantial representation of Italian might indicate an analysis targeted towards Italian speaking communities or regional events. Given its presence further exploration into Russian speaking networks could provide insights into online discussions. The smaller percentage allocated to languages may represent niches or smaller communities within the dataset. Ultimately understanding the language distribution helps shape approaches by focusing on each languages nuances and uncovering context insights that can influence observed interactions and behaviors, within the data.

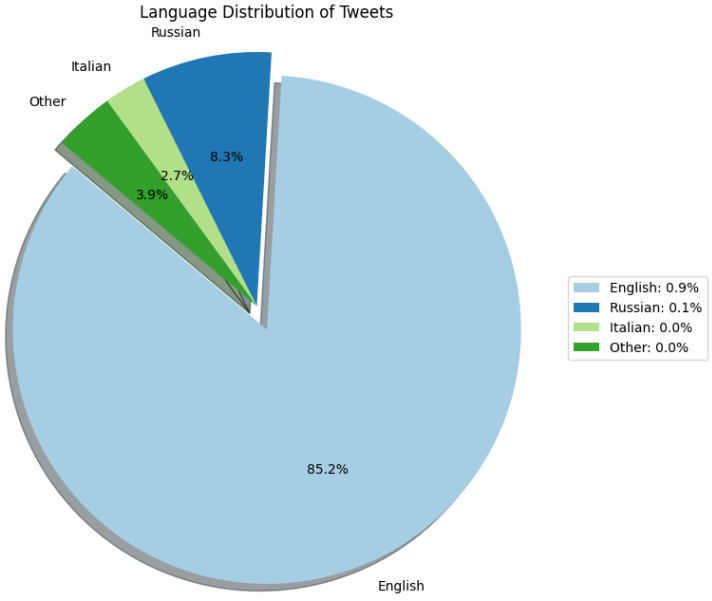

Figure 4.1: Language Distribution

4.2.2 Topics' extraction

The use of KNIME to apply Latent Dirichlet Allocation in topic modeling offers a valuable approach for understanding and organizing large amounts of unstructured text data, such as tweets. KNIME provides a reliable and user-friendly platform that enables users without coding expertise to perform complex data analytics tasks. By using its visual workflow functionality, one can easily implement LDA to categorize tweets into various topics based on word frequencies and patterns of occurrence, with the goal of identifying recurring themes or subjects. The utilization of latent Dirichlet allocation effectively condenses large text corpora into coherent topics, allowing for a better understanding of recurring themes. The probabilistic nature of LDA aids in assigning tweets to specific topics based on calculated probabilities, providing a nuanced approach to comprehending textual data.

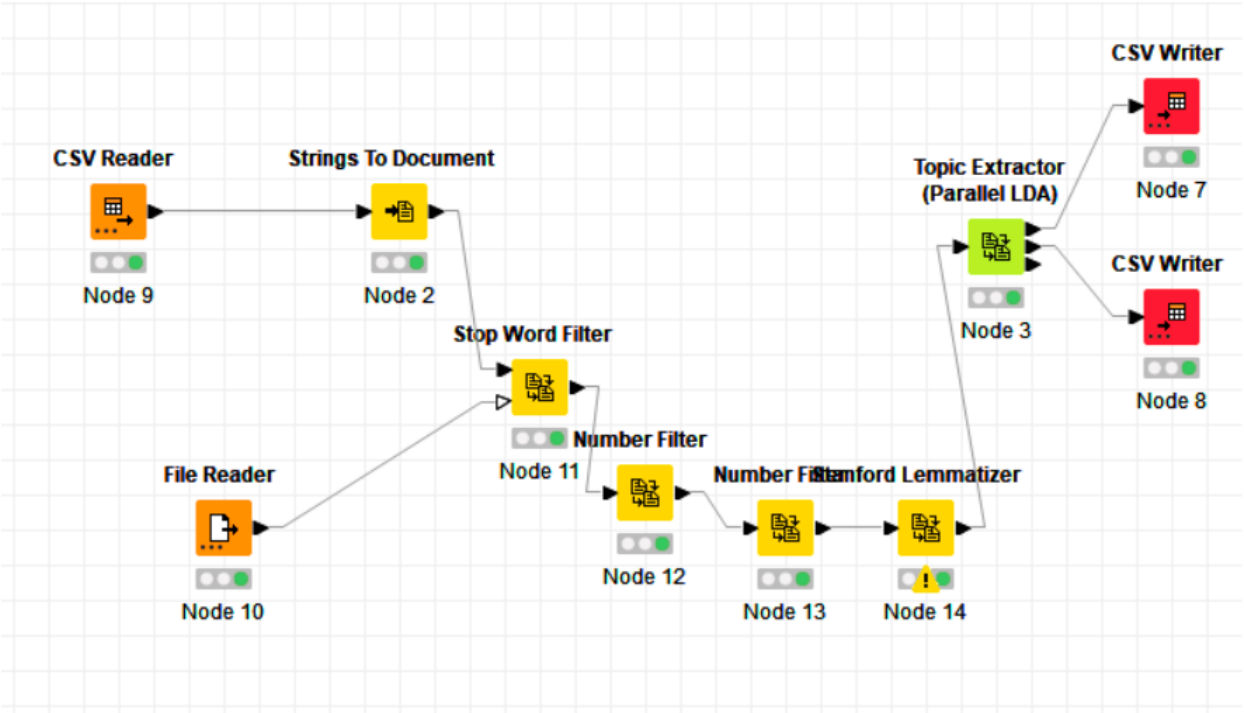

Figure 4.2: LDA on Knime

Topic	Words in the Topic
topic_0	mt, america, god, people, time, thank, stand, country, constitution, support, hear, love, life, gun, president, govt, real, military, nation, change
topic_1	watch, video, please, listen, president, check, interview, trump, share, news, support, live, music, radio, rally, artist, fuck, official, retweet, youre
topic_2	vote, mt, people, trump, read, tweet, speech, stop, american, congress, money, americans, looks, retweet, west, cruz, care, support, help, america
topic_3	house, day, win, story, history, black, im, please, donate, congrats, white, home, yall, love, people, truth, america
topic_4	tell, true, real, hey, thats, live, love, day, little, da, look, trump, truth, time, believe, feat, war, people, ill, world
topic_5	im, happy, love, black, prod, thanksgiving, day, check, lives, music, taking, matter, baby, brother, thank, dj, family, word, alert, fans
topic_6	cnn, cern, dig, artificial, wonderland, intelligence, bench, ordnung, bringt, stevens, olu, ein, ist, rza, er, mann, richtiger, frenemies, idols
topic_7	cnn, planes, flotus, claims, fly, separate, threat, difficult, chains, revere, voltaire, fools, vous, pense, Ã, nellz, monmouth, maleliens, dirtybird, syph
topic_8	coming, cant, network, soon, black, yes, whos, lost, radio, step, wall, talk, watch, youre, straight, ready, job, campaign, joining, media
topic_9	hillary, clinton, trump, obama, media, campaign, remember, news, red, ladies, email, emails, white, russia, liberal, fbi, fake, wikileaks, house, looking
topic_10	party, trump, check, book, donald, hosted, weeks, askem, heres, cartoon, massive, song, called, wait, talks, bernie, awesome, media, morning, app
topic_11	follow, join, game, twitter, live, tonight, play, thanks, check, miss, tune, guys, â9e, retweet, ?, fun, hashtag, hosted, oh, hey
topic_12	dani, bostick, impossible, park, votes, knocked, lot, gtgtgt, beach, ltltlt, words, socialism, damn, rinos, star, figure, cull, trina, thousand, worth
topic_13	thanks, daily, top, cruz, mt, trending, week, hashtag, list, ted, click, gt, support, obama, friends, thank, activist, clean, free, mixtape
topic_14	mt, killed, black, death, shot, police, threat, help, calls, video, political, people, obamacare, stop, america, perform, job, time, obama, cops

Table 4.2: Topics and Associated Words

Utilizing Latent Dirichlet Allocation (LDA) to uncover underlying patterns within the dataset, several pertinent topics have emerged, each shedding light on the potential roles social bots might play in online discussions. Beginning with the theme of "Political Engagement and National Identity," characterized by terms like "america," "constitution," "president," and "nation," there is a discernible focus on political dialogues and the cultivation of national identity. Within this context, social bots could potentially amplify political messages, influence public perspectives, and foster a sense of patriotism.

The subsequent theme, "Multimedia Sharing and Political Figures," stands out with terms such as "watch," "video," "trump," and "news." This indicates a noteworthy emphasis on the dissemination of multimedia content tied to political figures. Social bots may contribute by circulating videos, interviews, and news clips to propagate specific political narratives, thereby influencing the portrayal of these figures in the public eye. Moving forward, the theme of "Voting and Political Action" revolves around keywords like "vote," "trump," "congress," and "support." This theme suggests discussions that center on political participation and voting behaviors. Social bots could potentially engage in conversations that encourage support for particular candidates, shape discussions about voting patterns, and potentially impact electoral outcomes.

Exploring historical and cultural contexts, the fourth theme encompasses terms like "history," "black," "america," and "white." These terms suggest conversations that delve into historical and cultural dimensions. In these discussions, social bots might mirror societal divisions, historical events, and cultural nuances, which could potentially sway public perceptions on these matters.

The theme of "Truth and Authenticity" comes into focus with words like "true," "real," and "truth." These terms highlight conversations that delve into authenticity and misinformation. This theme raises the possibility of social bots being involved in disseminating or countering misleading information, potentially influencing public perceptions and contributing to the ongoing discourse on information veracity.

Transitioning to a more positive realm, the sixth theme underscores terms like "happy," "love," and "family," suggesting discussions that foster positive emotions and community engagement. Social bots might contribute to dialogues that cultivate positivity and strengthen communal bonds among users.

Within the scope of technology and artificial intelligence, the seventh theme, defined by keywords like "cnn," "artificial intelligence," and "intelligence," points to discussions concerning the fusion of technology and AI. Social bots may engage in these conversations, contributing to narratives on technological advancements and the societal implications of AI. The theme

of "Conspiracy and Skepticism" takes center stage with terms like "claims," "revere," and "fools." These terms hint at discussions centered around conspiracy theories and skepticism. Social bots could potentially amplify these unconventional concepts, potentially exacerbating polarization and fostering distrust.

Shifting gears, the theme of "Anticipation and Engagement" encompasses words like "coming," "soon," and "ready," suggesting conversations marked by anticipation and active involvement. Social bots might contribute to generating excitement around upcoming events or initiatives, influencing the level of interest and attention within the discourse.

The theme of "Party Affiliations and Political Commentary" is characterized by terms such as "party," "trump," "book," and "donald." This theme suggests discussions that revolve around party affiliations and political commentary. Social bots could engage in conversations about political figures, hostings, and various perspectives, potentially shaping the narratives surrounding political parties and their stances.

The theme of "Social Media Engagement and Online Activities" stands out with terms like "follow," "join," "game," and "twitter." This theme indicates a focus on social media engagement and online activities. Social bots might contribute to discussions about following, live events, hashtags, and interactive engagements, potentially impacting user behavior and participation.

The theme of "Individual Expression and Cultural Influences" encompasses terms like "im", "happy", "love", and "black". This theme suggests conversations that highlight individual expression and cultural influences. Social bots could contribute to dialogues that celebrate individual expressions, address cultural matters, and engage with music, potentially fostering a sense of cultural identity and connection.

Within the theme of "Technological Advancements and AI Discourse," keywords like "cnn", "cern" and "artificial intelligence" come to the forefront. This theme suggests discussions centered around technological advancements and AI. Social bots may engage in conversations about AI, artificial intelligence, and innovative developments, potentially shaping narratives about the impact of technology on society.

The theme of "Public Perception and Political Controversies" features keywords such as "hillary," "clinton," "trump," and "wikileaks," signaling discussions focused on public perceptions and political controversies. In these conversations, social bots could contribute to discussions about scandals, emails, and political maneuvers, potentially shaping the narratives surrounding these contentious topics.

In total, there are 15 distinct themes identified in the topics extracted from the data. Each theme represents a cluster of related terms and discussions that social bots are potentially

engaging with, contributing to a diverse landscape of online conversations and discourse

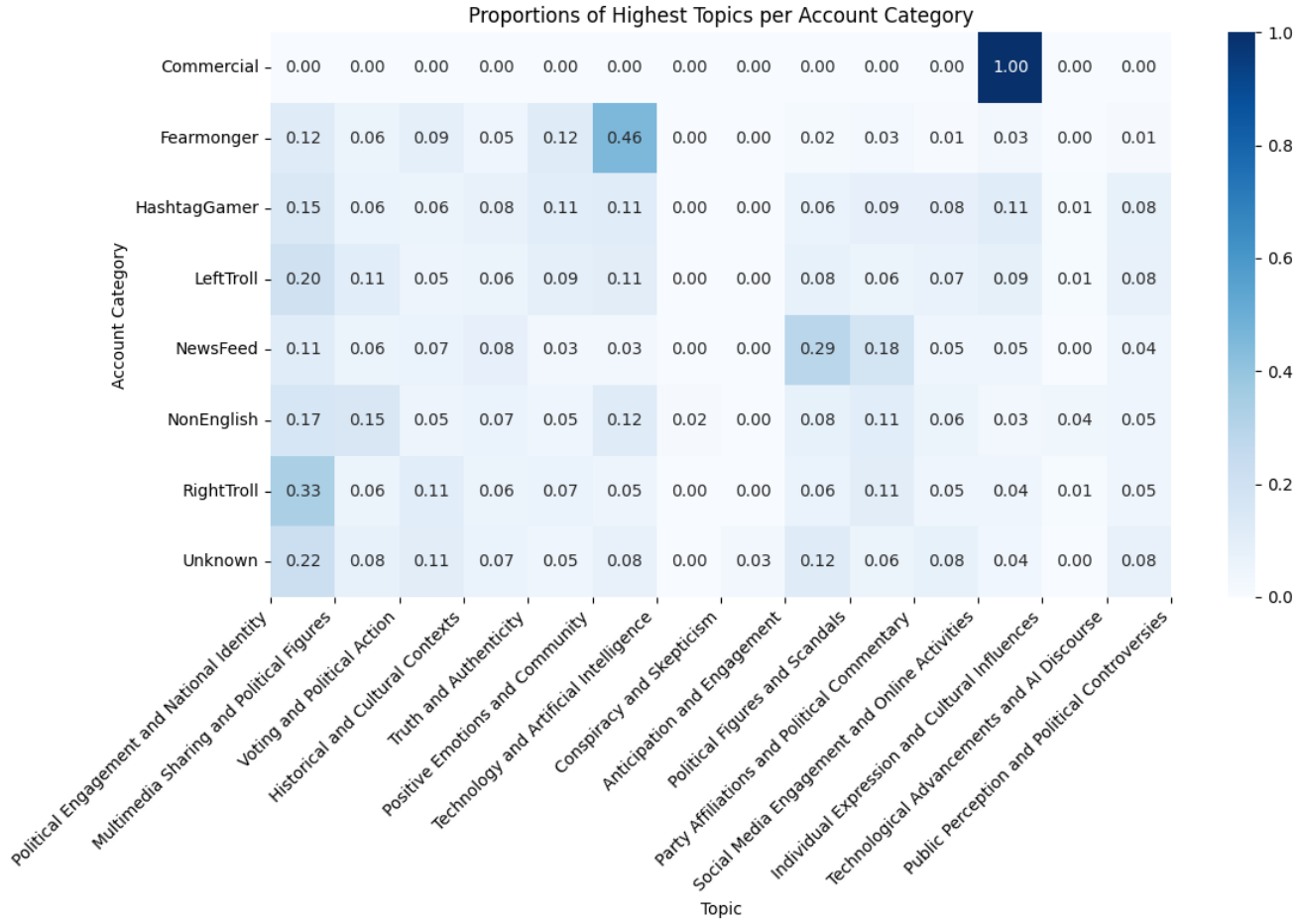

Figure 4.3: Proportion of highest topics per account category

Digital conversations are a complex tapestry formed by different types of accounts, each playing a role in shaping how we interact online. The following investigates the distribution of tweets across various account categories, revealing valuable insights into the intricate dynamics that characterize this digital landscape.

Within this context, a significant portion of tweets is attributed to "RightTroll" accounts, accounting for about 55.34% of the total. This prominence serves as evidence of the considerable impact generated by profiles promoting right-leaning populist narratives. The connection between these accounts and Donald Trump's presidency underscores the dynamic nature of online political discourse in contemporary times.

In a notable contrast, the "LeftTroll" accounts, which make up approximately 26.26% of tweets, stress the importance of social liberalism intertwined with discussions on culture and

racial identity. They actively promote hashtags such as blacklivesmatter, showcasing the influential role of social movements in shaping online conversations and impacting public discourse.

However, the analysis expands beyond political conversations. The emergence of "HashtagGamer" trolls, who account for 9.62% of tweets, adds an unconventional element to the discourse. These trolls engage in word games on Twitter and exhibit a sophisticated ability to construct narratives that incorporate themes from both "RightTroll" and "LeftTroll" categories. Interestingly, accounts categorized as "NonEnglish," which make up 8.54% of tweets, transcend language barriers and actively participate in shaping global discussions. Even the small presence of "Fearmonger" trolls, contributing only 0.17% of tweets, carries significance as they have the potential to spread misinformation and instill fear- most notably seen with the contaminated turkey narrative. Their adaptability across different troll categories highlights their role in fostering discord within discourse communities

The influence of right-leaning, left-leaning, cultural, and fear-inducing messages on public opinion and perception is evident. This analysis provides a comprehensive examination of thematic content and account categories, offering a detailed insight into the complex dynamics that shape digital communication in various contexts.

Upon unraveling the intricate categories, a comprehensive narrative unfolds and sheds light on the interdependent relationship between account categories and discourse themes. The cohesive alignment of tweet distribution and retweet patterns serves as evidence for the influential role that certain categories play in shaping discussions through diverse strategies. This collective understanding presents an opportunity to delve deeper into exploring the motivations, impacts, and tactics employed by these distinct categories, ultimately enhancing our understanding of the rapidly changing realm of digital communication.

Furthermore, the examination of high-retweet account categories offers profound insights into their content, engagement patterns, audience composition, and prevalent storylines. This exploration unveils the influential impact of these accounts on shaping digital discussions, characterized by recurring themes, language patterns, and emotionally resonant expressions that profoundly connect with their followers.

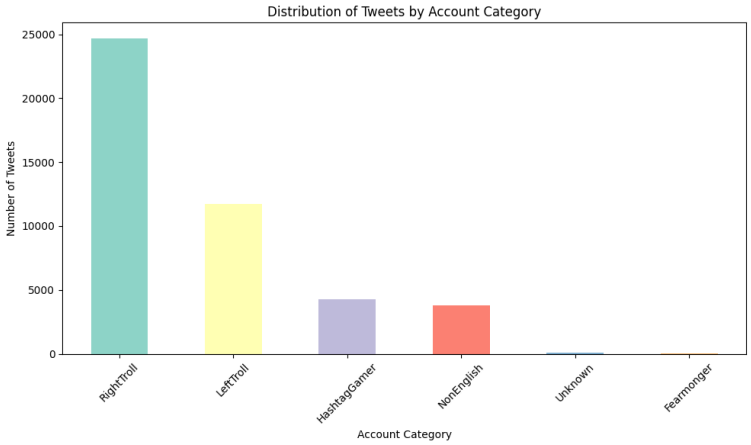

Figure 4.4: Distribution of Tweets by account category

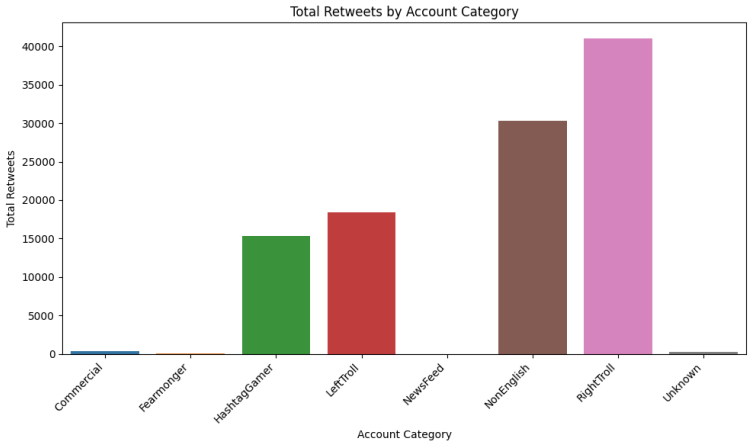

Figure 4.5: Distribution of retweets by account category

4.2.3 Comparative Data Framework: An Integration of Multiple Twitter Datasets for Robust Disinformation Detection

To broaden the comprehensiveness and credibility of this study,supplementary datasets have been included for comparative analysis. One such dataset is the Sentiment140 dataset, obtained prior to 2009 when Russian troll activity was presumably less prevalent. Even though this dataset is not heavily influenced by disinformation campaigns it's important to acknowledge that some adjustments were made, such as removing emojis. This might affect its comparability, with the Troll Tweets dataset. The second dataset comprises tweets from

celebrities providing a contrasting sample in terms of demographics and context. By including tweets it's possible to thoroughly analyze syntactic features. However it's important to note that due to the demographics and inherent biases in celebrity communication certain variables may be introduced that are less relevant for detecting disinformation but still significant for analysis. Combining these datasets allows for an examination that helps identify both general and specific characteristics of disinformation.

By comparing attributes across data samples this study aims to identify indicators while evaluating the models performance, in different scenarios. This improved approach enhances. Applicability of the findings deepening the understanding of how contemporary disinformation campaigns operate. The selected feature set has been thoughtfully designed to achieve two objectives.

The primary goal is to identify the techniques employed by individuals who spread misleading information, known as "Amplification Metrics." These metrics include tracking URL counts, hashtag counts, and mentions of Twitter handles, all aimed at promoting widespread message dissemination. After achieving high accuracy in classifying these agents using the existing features, it was found unnecessary to include emotional arousal indicators like exclamation marks. Additionally, this study reveals that these agents often adhere to specific guidelines regarding message length, indicating operational constraints they operate within.

The second category, refferred to as "Linguistic Inconsistencies," targets non-native English speakers, particularly Russian trolls, by taking advantage of their distinctive linguistic characteristics. The identification of anomalies, between the structures of Russian and English is achieved by observing the differences in comma and dash usage. While advanced techniques in natural language processing can be explored to detect linguistic errors the current set of features is enough to achieve a high level of accuracy in detection. The effectiveness of these selected features is further supported by data visualization. Histogram analysis shows that while there are some overlapping distributions between trolls and non trolls noticeable differences also emerge. Notably there are variations in the length of tweets and maximum word length between these two groups. It's also worth mentioning that emojis are used at frequencies in both datasets despite one being pre cleaned. This suggests that regular social media users tend to use emojis compared to troll posts where emojis are noticeably absent.

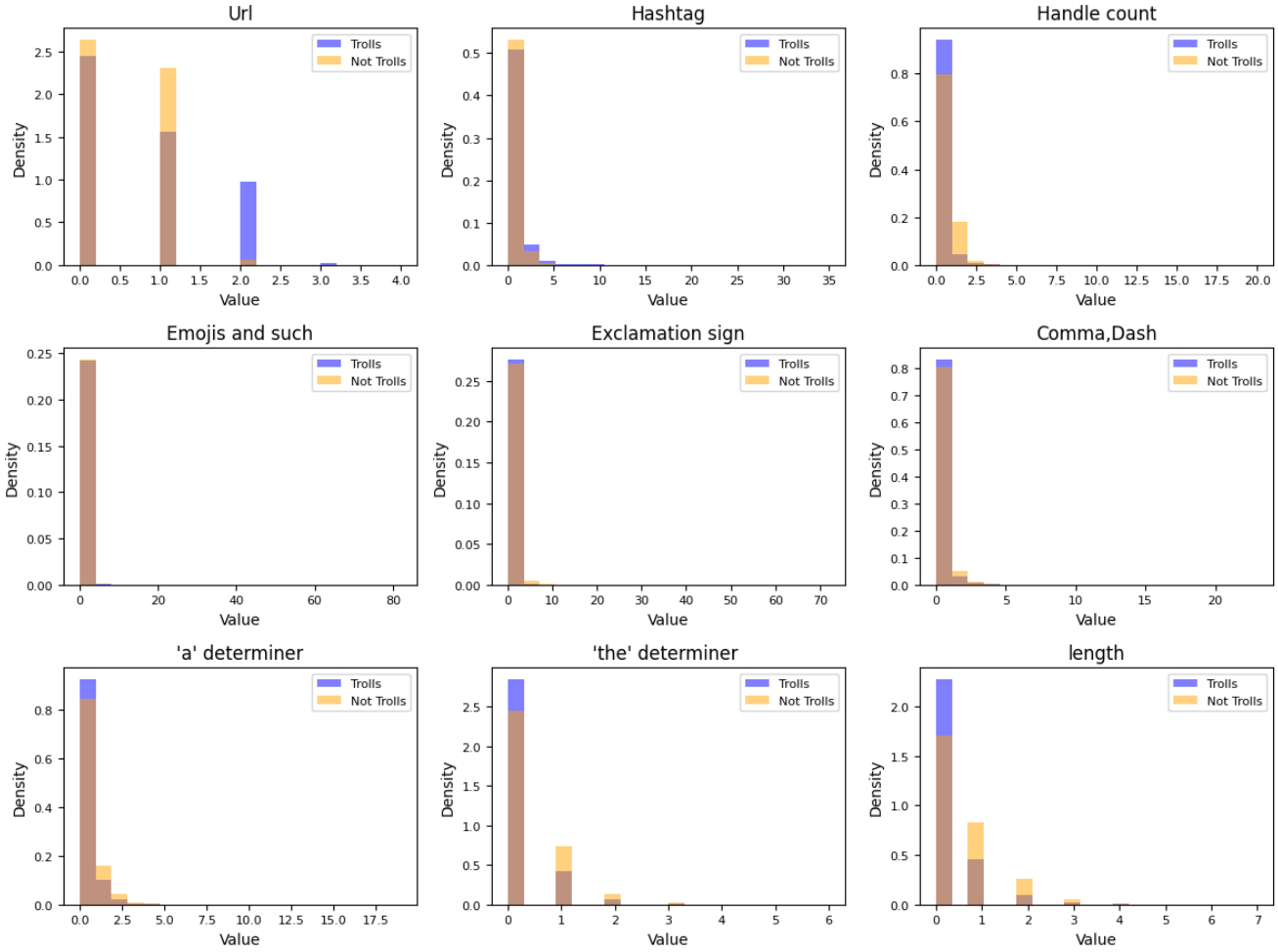

Figure 4.6: Trolls vs. Non-Trolls features

Based on the analysis it seems that agents involved in disinformation campaigns have quotas that limit their activities. This limitation leads to use of tweets with lengths. Additionally these agents often include words, in their tweets to incorporate certain keywords or phrases based on their operational guidelines. Consequently, this thoughtfully selected set of features not only enables effective identification of trolls but also offers valuable insights into the coordinated tactics employed in widespread disinformation campaigns.

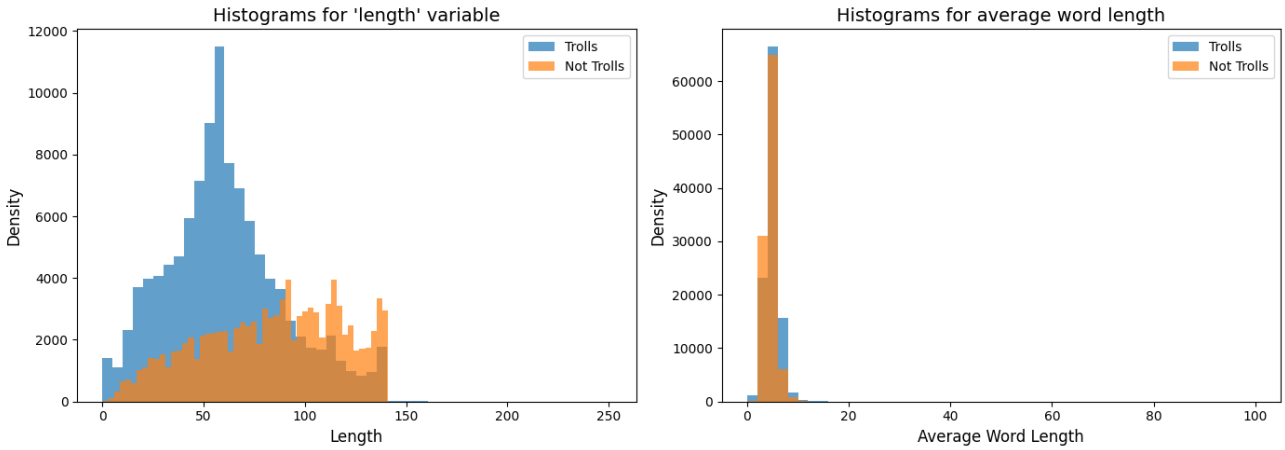

Figure 4.7: Trolls vs. Non-Trolls features

4.3 Temporal Analysis

Gaining insights into the temporal aspects of online interactions is crucial in comprehending the intricate relationship between social bots and digital discourse. This stage of analysis seeks to reveal the intricacies of engagement patterns over time, providing valuable information on trends, activity cycles, and the effects of significant events. Exploring the temporal dimensions of social bot activity allows to understand how these entities adjust their tactics, react to external stimuli, and shape discussions within digital platforms.

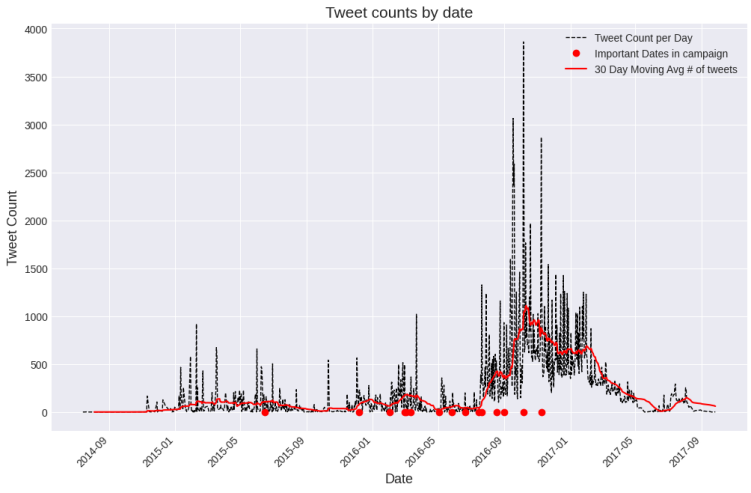

Figure 4.8: Tweets count by date

Date	Tweet Count	Percent Change (%)
2015-06-16	3	50.000
2015-12-07	219	204.167
2016-02-01	18	1700.000
2016-03-01	143	-71.053
2016-03-03	6	-92.105
2016-03-11	64	-69.524
2016-05-03	38	216.667
2016-05-26	6	-50.000
2016-06-20	201	1156.250
2016-07-15	47	17.500
2016-07-21	1327	349.831
2016-08-17	534	20.270
2016-09-01	337	-63.919
2016-10-07	2222	-42.450
2016-11-08	2867	145.043

Table 4.3: Tweet Counts and Percent Changes

Notably, on various dates, there were significant shifts in tweet counts, some experiencing notable increases of up to 350%. Equally important are instances of negative changes, highlighting the fluctuating nature of digital engagement. It's worth clarifying that these dates, though not presented in chronological order, mark key events rather than singularly focusing on elections. This distinction helps sidestep potential political associations, keeping the analysis neutral. The "percent change" metric, calculated in relation to the previous day, adds depth to the understanding of tweet activity patterns. The data points to a range of

events that drew considerable attention, reflecting the influence of temporal factors on online conversations. These findings resonate with the evolving nature of digital engagement, where events can trigger both surges and declines in social media activity. This research enhances our grasp of the interconnectedness between real-world events and their digital amplification, offering valuable insights for a broader comprehension of information dissemination dynamics.

4.4 Network Analysis

4.4.1 User Metrics

One noteworthy aspect pertains to the asymmetry of influence distribution between users categorized as 'popular' and 'unpopular,' based on a well-defined metric of follower-to-friend ratio. Contrary to what's found in literature, which often emphasizes the influential role of popular users in spreading information this study reveals a more intricate scenario. 3% of the tweets in the dataset were attributed to users yet these tweets held a considerable impact of 21.8% overall. This suggests that these users function as amplifiers likely due to their follower networks that enhance the visibility and reach of their tweets. Given the susceptibility of social media platforms to manipulation by actors such as bots and troll accounts it is vital to comprehend the influence wielded by these users in shaping public conversations in order to combat misinformation. On the other hand a significant 78.18% of tweet influence was traced back to users categorized as unpopular (follower-to-friend ratio < 2), who were also responsible for 87% of the total tweets. Although these users individually may not have extensive reach, their collective influence should not be underestimated, particularly in the realm of social bots and disinformation campaigns. These 'unpopular' users could potentially form a network of bots or manipulated accounts that work together to influence or distort the flow of information. Thus, while the role of popular users as key nodes in information dissemination is unquestionably important, this analysis underscores the equally critical role that a large mass of less. Building upon the analysis there is added complexity in examining patterns of when these accounts were created. Notably there was an increase in the creation of "popular" accounts during 2014 particularly concentrated in May and June. The timing of this clustering raises questions about coordination, behind the emergence of these influential accounts. Considering how social networks operate and the gradual process typically required to accrue influence, it is plausible to conjecture that these accounts were strategically positioned for long-term follower growth and an amplified scope of impact over

time.

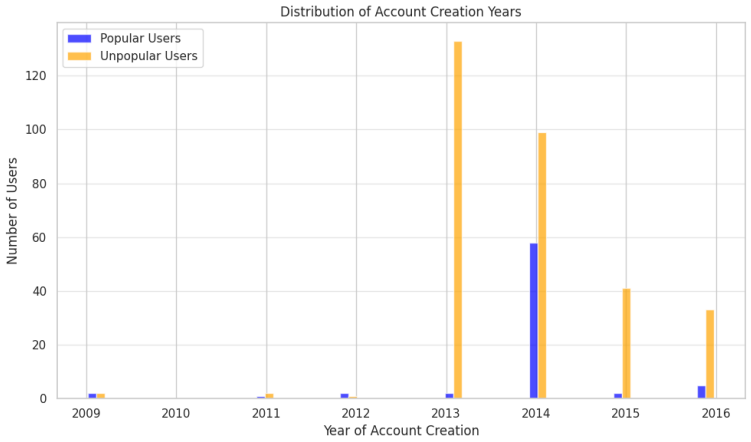

Figure 4.9: Distribution of account creation years

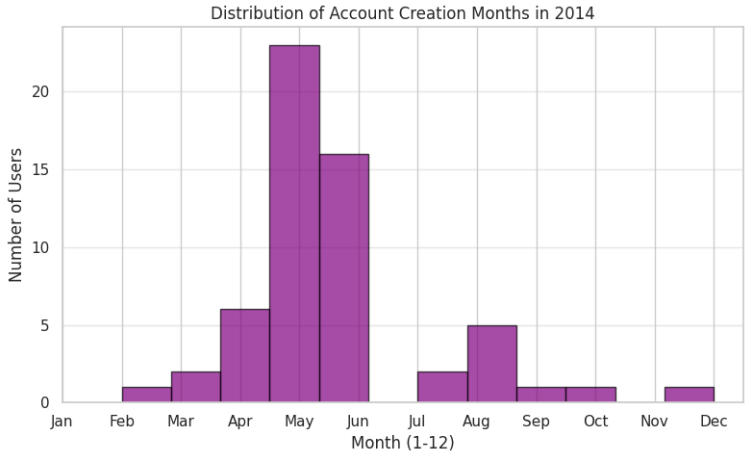

Figure 4.10: Distribution of account creation months in 2014

Interestingly, the 'less popular' accounts also show temporal clustering but are more scattered, with significant spikes in August 2013 and again in May and June 2014. The similar time periods of account creation for both 'popular' and 'less popular' accounts could suggest coordinated efforts to create these accounts, potentially aiming to disseminate information or disinformation effectively. While the 'popular' accounts act as amplifiers due to their wide reach, the 'less popular' ones can function as accomplices in influencing public opinion by creating the appearance of widespread agreement or adding noise to undermine genuine discussions.

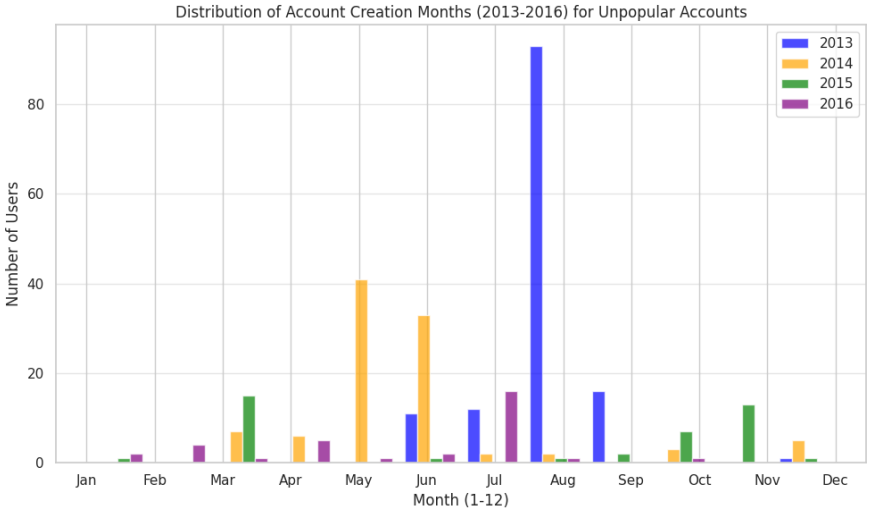

Figure 4.11: Distribution of account creation months (2013-2016) for Unpopular accounts

This trend is especially important for experts in the field of social bots and disinformation studies. For example, the clustering of activity over time could indicate a coordinated campaign, which may be either naturally occurring or orchestrated for manipulative purposes. The synchronization in the creation periods of both "popular" and "unpopular" accounts suggests a multi-tiered strategy, where popular accounts are used to spread messages widely while unpopular accounts operate on smaller scales. These accounts may work together to perform tasks such as reinforcing messages or generating rebuttals.

4.4.2 Network Dynamics and Dissemination Strategies

The analysis uncovered that retweets accounted for 73% of the activity, suggesting a primary approach centered around amplification rather than generating original content. It is worth noting that only 2% of these retweets amplified content produced by other Russian troll accounts within the same network, indicating an intricate strategy that leverages existing material.

There were disparities identified between the most active users on Twitter and those who received the highest number of retweets within the troll network. Users like AmelieBaldwin and hyddrox solely focused on sharing content from other accounts, whereas TEN_GOP and ChrixMorgan generated a substantial amount of their own original posts that were frequently shared by others. This implies an allocation of roles, with certain accounts serving as amplifiers while others fulfill the role of creators.

Further analysis uncovered that the number of retweets doesn't necessarily indicate pop-

ularity in this network indicating an coordinated operational approach. Additionally the retweeted external accounts displayed varying ideological orientations, which adds complexity to the spread of false information. To effectively combat disinformation it is crucial to adopt a strategy that goes beyond content moderation. Factors such as network dynamics and behavioral tactics must be taken into account to address this issue successfully. The analysis below highlights patterns in user behavior that have implications, for understanding how information and disinformation are disseminated. Upon examination of the accompanying figures it becomes clear that three distinct tiers can be observed.

4.4.3 Stratified Architectures of Information and Disinformation

In the following cluster map, each row represents a unique user, and each column represents an hour of the day. The color intensity indicates the frequency of tweeting: darker colors mean more tweets, and lighter colors mean fewer.

The clustermap provided reveals that with the exception of a percentage of users who tweet often there isn't a significant preference for specific time periods to share tweets. This finding aligns with studies that suggest frequent tweeters tend to distribute their tweets throughout different times. The neutral pattern of tweet activity suggests that either an automated system is responsible or there is a group of operators spread across geographic locations reducing the influence of time zones on tweet frequency. It also raises the possibility that high volume tweeters may not strategically select optimal posting times to maximize engagement, which is typically considered important in information operations. All these consistent findings at different stages strengthen the reliability and robustness of the observed patterns.

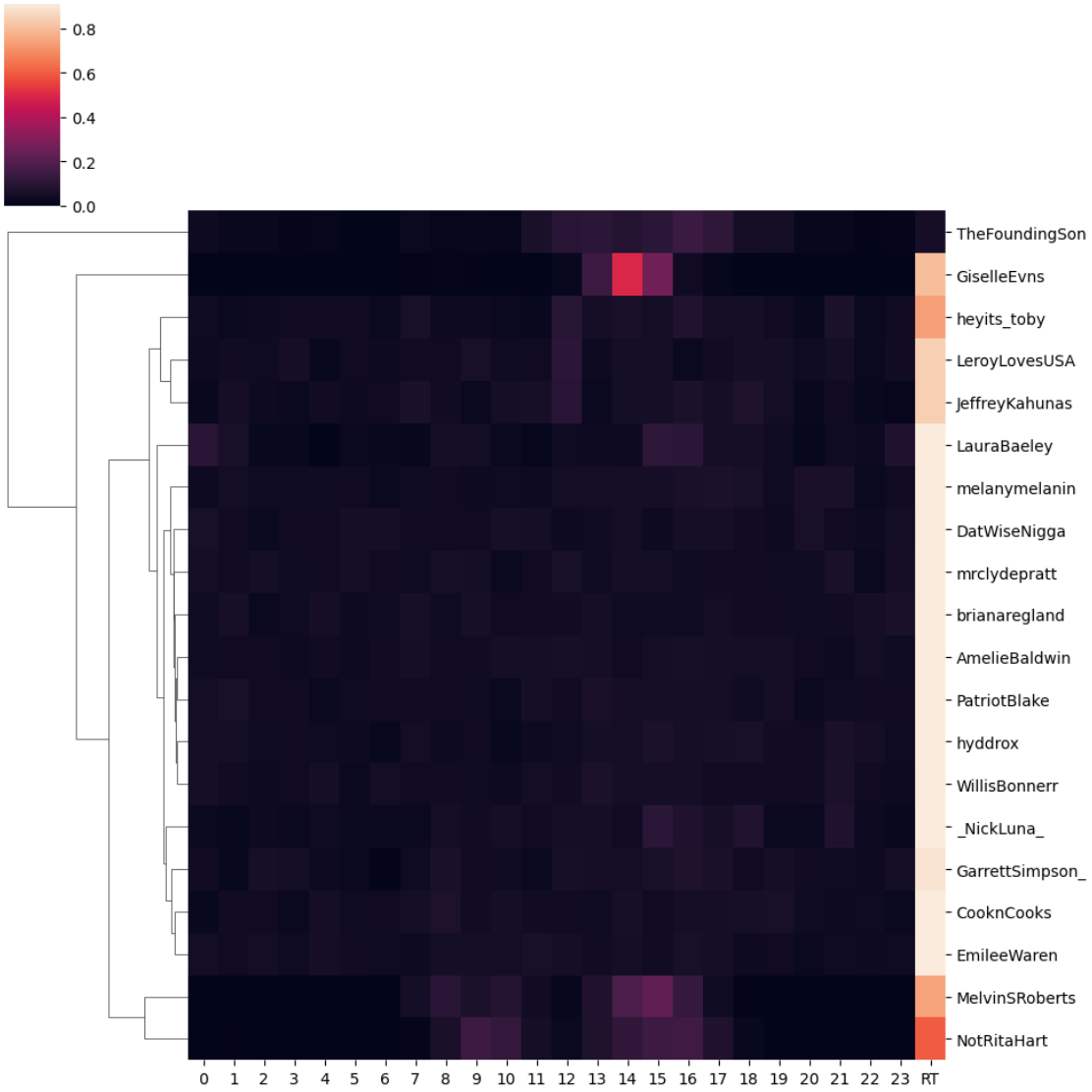

Figure 4.12: Tweeting pattern over a 24 hour period for the top 20 tweeters

In a detailed analysis of tweeting patterns, there is an intriguing contrast between the top 20 users who tweet most frequently and the top 20 users who are retweeted the most. The latter group demonstrates more organized temporal behaviors, indicating that their tweeting schedule is well-planned. In particular, distinct clusters of similar tweeting behavior can be observed within this group. Usernames such as WorldOfHashtags, DanaGeezus, ChrixMorgan, and GiselleEvns show a concentrated burst of activity from 1-3 PM. On the other hand, users like Pamela_Moore13,Crystal1Johnson,TodayPittsburgh,TEN_GOP,and tpartynews have

a consistent distribution of tweets throughout the evening hours from6PM to2AM

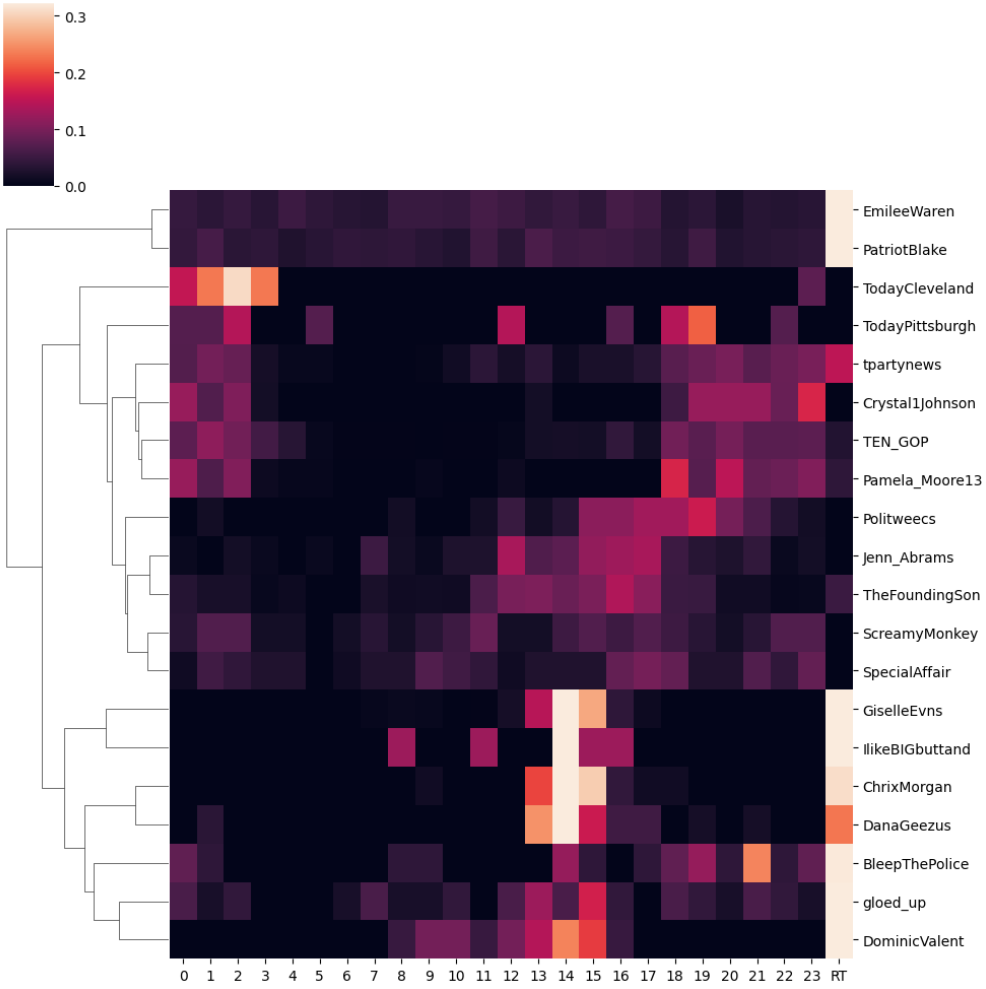

Figure 4.13: 24 hour period for the top 20 most retweeted users

Following the analysis of the data on tweet patterns and content dissemination dynamics, distinct categories of actors emerge within the social media information ecosystem. The "Content Initiators" stand out as primary sources of original content. They seem to be predominantly operated by humans who contribute original content to the ecosystem. These accounts exhibit activity patterns that mirror natural human rhythms and traditional news cycles, displaying periods of inactivity as well as peak hours. They act as the foundation for spreading information throughout the network, wherein messages are selected for ampli-

fication. Another important point to consider is the presence of "Amplification Bots" on social media platforms. These bots operate with remarkable efficiency, strategically retweeting content in order to maximize its virality. This suggests that they are driven by algorithms designed to take advantage of the platform's mechanics for increased exposure. As a result, these bots play a pivotal role in disseminating both accurate information and misinformation to larger audiences. Furthermore, in the realm of content dissemination, there exists a group of "Aggregator Influencers." These individuals or semi-automated accounts serve as important intermediaries between creators of original content and a wider audience. By consolidating information from various sources into concise messages, accounts like 'GiselleEvns' play a crucial role in amplifying messages.

This dynamic structure fosters an intricate yet efficient system for message propagation. The process involves "Content Initiators" being retweeted by "Aggregator Influencers," who are further amplified by "Amplification Bots." This hierarchical framework of transmitting information displays distinct patterns over time. While original content generators tend to follow human-like schedules with identifiable periods of decreased activity, bots exhibit more structured and predictable patterns aimed at maximizing their reach.

The model is highly valuable for comprehending the structure of disinformation campaigns. It effectively identifies the major actors and behaviors within the network, offering insights into strategies utilized to enhance message visibility, regardless of their intent being benign or malicious. This comprehensive amplification model presents a sturdy foundation for future empirical investigations. The incorporation of advanced machine learning techniques such as clustering and social network analysis could reinforce these findings, which are crucial in developing effective counter-disinformation strategies.

4.4.4 Information Cascades and Amplification in Russian Twitter Networks

The emphasis of this section is on identifying tweets that have a high number of retweets, specifically focusing on the top 200 tweets. These selected tweets will be further analyzed in terms of their temporal patterns. Additionally, the time difference between the original tweet and its corresponding retweet will be calculated to gain a more comprehensive understanding of how information spreads on this platform.

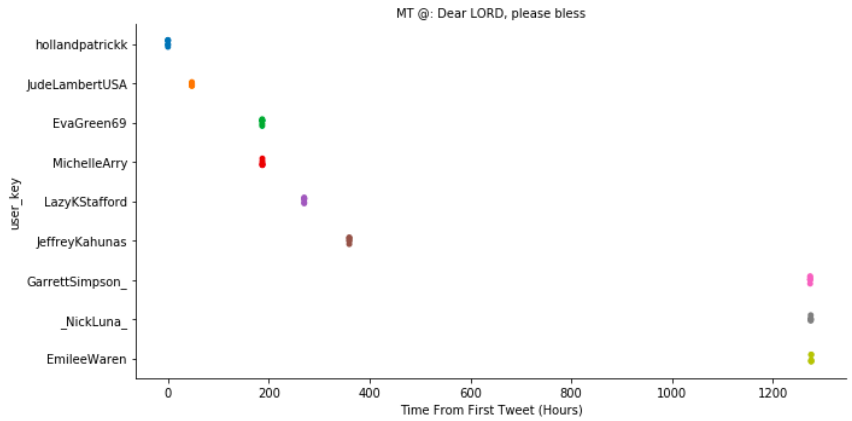

Figure 4.14: Linkage between users

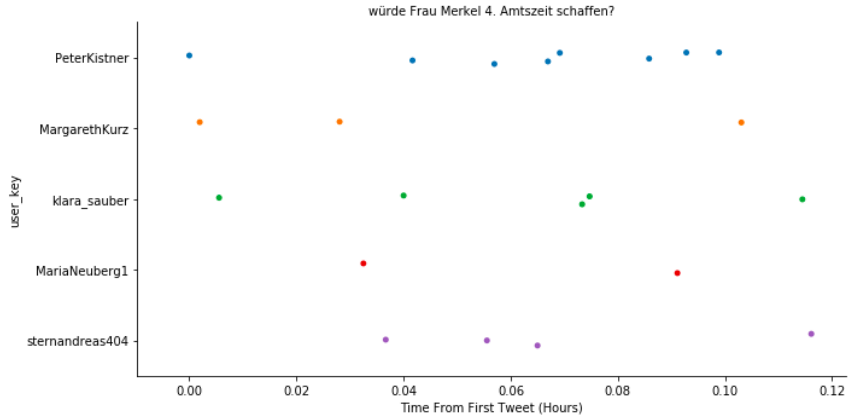

Figure 4.15: Linkage between users

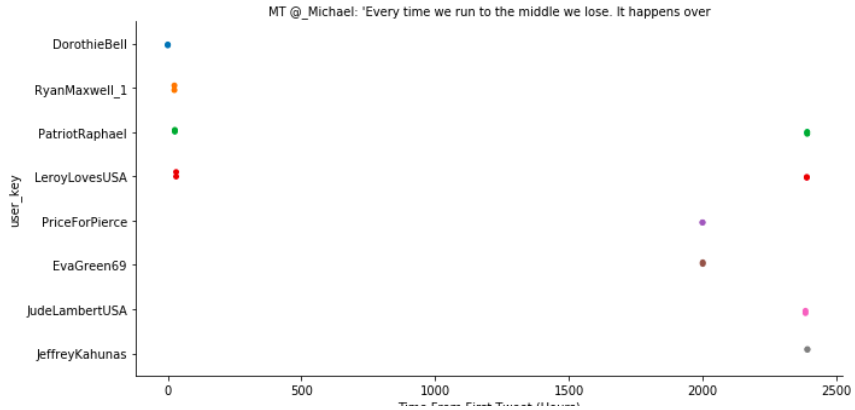

Figure 4.16: Linkage between users

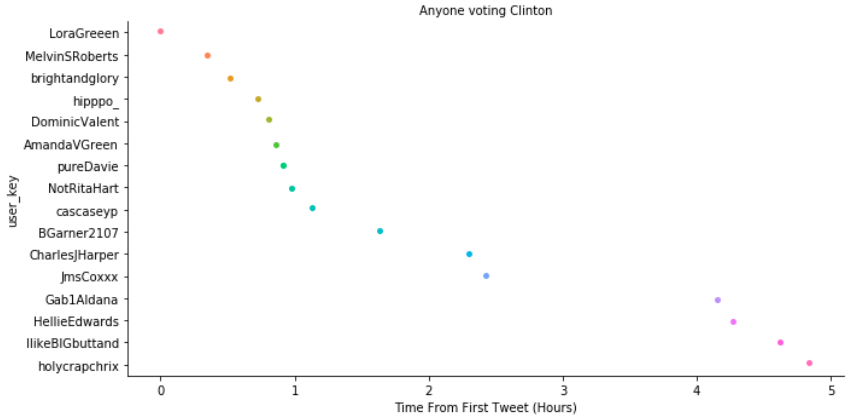

Figure 4.17: Linkage between users

The graphs reveal interesting findings about retweet behavior on social media platforms. Through data analysis, two distinct temporal patterns of retweeting can be observed. The first pattern, referred to as "Reciprocal Retweeting," involves a limited group of users sharing either their own content or that of others within the same network over an extended period, sometimes up to a year (e.g. Figure 4.14). This behavior indicates the presence of insulated communities where specialized content circulates and resonates among like-minded individuals, whether it be related to ideology, commerce, or any other subject matter.
On the other hand, a different pattern that we will refer to as "Retweet Chains" shows a sequence of interconnected retweets involving various users (Figure 4.17). Each tweet in this chain is shared by a distinct user, who is then retweeted by another unique user, resulting in a cascade of information sharing. Interestingly, these chains occur within narrow time frames, ranging from as short as 20 minutes to around 5 hours. This rapid succession suggests strategic amplification methods driven by algorithms that aim to exploit platform-specific dynamics like trending topics or fast-paced news cycles.
It is essential to conduct a detailed investigation into the different temporal behaviors exhibited by these patterns. This will help determine any correlations with specific categories of content such as misinformation, political propaganda, or commercial messaging. A thorough understanding of these temporal patterns plays a crucial role in comprehending how information spreads, both accurately and misleadingly, on a broader scale.

In order to examine the patterns of retweet sequences and determine if certain users tend to appear before others, a chronological representation is used on the X-axis. This axis indicates the elapsed time from when the original tweet was posted. On the other

hand, unique usernames are categorized on the Y-axis as identifiers for individuals who have engaged with the original content through retweets.

To enhance the analysis, categorizing the origin of each retweet based on color differentiation provides a more nuanced understanding of the retweet chain. By examining the density and dispersion along the X-axis, viewers can assess the speed at which the retweet chain unfolds over time. Dense clusters suggest quick dissemination within a specific timeframe, while scattered data points indicate a slower diffusion pattern.

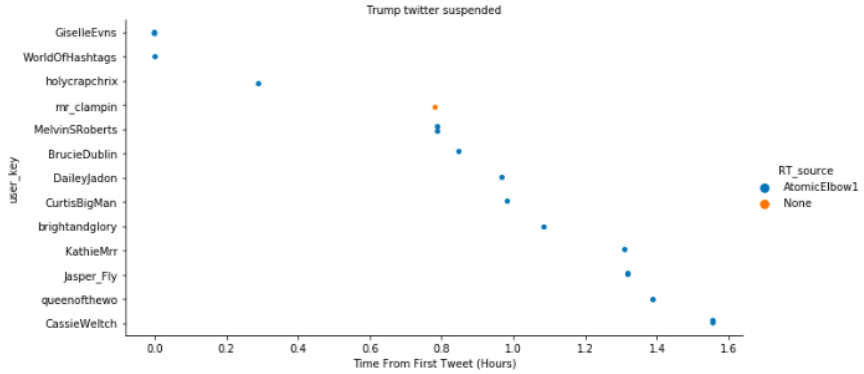

Figure 4.18: Chain Retweets

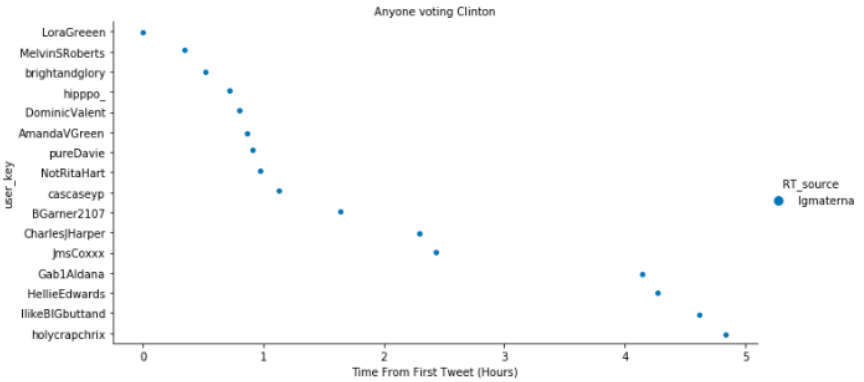

Figure 4.19: Chain Retweets

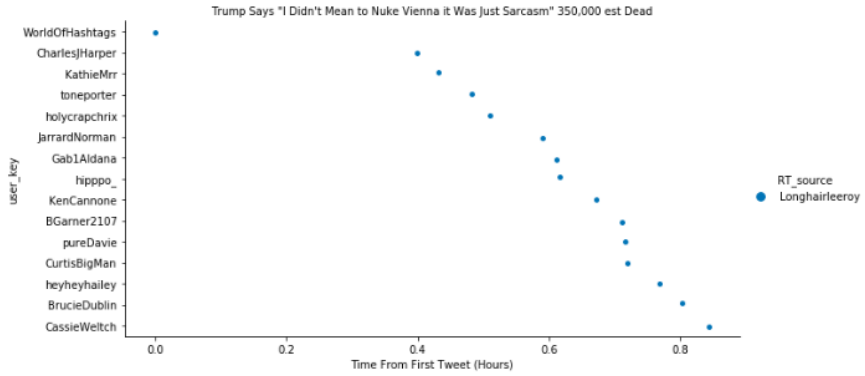

Figure 4.20: Chain Retweets

Furthermore, the initiators in the Sequential Cascades are rarely the original creators of the messages. Their strategic placement within the sequence leads to two noteworthy observations: 1) The dispersion of the initial message among a larger audience, supporting the "diffusion of innovations" theory (Rogers, 2003), and 2) The lack of a predetermined retweet sequence, introducing an element of randomness that makes detection or prediction challenging.

This particular pattern reveals a centralized mode of information propagation where one user's content becomes the focal point for further dissemination. It is crucial for observers to note that, in this ripple model, the timeline of retweets along the X-axis essentially maps the velocity of each concentric ripple as it moves away from the initial source. The homogeneity of the original user being retweeted across these ripples highlights the gravitational pull of the original content, reinforcing its significance in the network dynamics. The intricate nature of information propagation strategies includes both human-operated and algorithmic accounts. These strategies not only circulate internal content but also import and amplify external material. This highlights the importance of trans-network counter-disinformation interventions, as well as the use of advanced graph-theoretic and machine learning techniques for detecting anomalies in real-time.

4.4.5 Network Dynamics and User Roles in Retweet Behavior

Focusing on the dynamics of media networks of individual tweets provides a comprehensive perspective on how information spreads. By analyzing the 10 retweeted users and the 20 users who retweet them that it's possible to pinpoint influential nodes in the network. This approach helps understand how information, including disinformation moves within ecosys-

tems.

By concentrating on these nodes it's possible to gain valuable insights into the mechanics of influence and digital engagement nuances. This focus allows to identify both accounts and those that use strategic methods to gain influence, which is crucial when examining how disinformation spreads. Analyzing networks in this way becomes a tool for uncovering vulnerabilities and robust features, within communication networks.

Therefore adopting a network based approach enhances our understanding of disinformation spread. Strengthens our ability to combat it. It also aids in formulating policy recommendations by identifying areas where targeted interventions can make a difference—an objective of our broader research.

Figure 4.21: Directed Graph of 10 most retweeted accounts

The directed graph analysis of retweet patterns uncovers distinct roles among users in propagating content from the top 10 most retweeted accounts. It is noteworthy that some users, such as MelvinSRoberts, emerge as key nodes in this network. Their high in-degree

centrality, represented by a medium-sized black dot in the visualization, implies they serve as major amplifiers of multiple top-retweeted users.

On the other end of the spectrum are users with specific allegiances to single sources. Users like mr_clampin and Jasper_Fly in the lower right quadrant of the graph are exclusively connected to DominicValent. Similarly, melanymelanin and BleepThePolice have unique affiliations with gloed_up. These users don't appear to diversify their retweeting behavior and instead focus on propagating content from one primary source.

These findings illuminate the different strategies and roles users adopt in the retweet network. While some act as broad amplifiers, enhancing the reach of multiple influential accounts, others operate in a more targeted manner. Understanding these nuances is crucial for anyone looking to comprehend the intricacies of information dissemination, including the potential for disinformation spread, on social media.

4.4.6 Inter-bot Interaction Analysis

Finally, to gain insights on how social bots operate and influence conversations, their interactions will be examined. To do this it's been introduced a column in the dataset, for categorization purposes. The values in this column are either 'true' or 'false'. A 'true' value indicates that a bot has interacted with another bot in a tweet, while a 'false' value suggests that the bot has mentioned an actual user account. This approach effectively separates interactions between bots from those involving users allowing us to develop a comprehensive understanding of bot behavior patterns within the data.

After implementing the column to differentiate between interactions involving bots and humans it was found the following; out of 1,153,605 instances analyzed there were cases where bots interacted with humans (marked as 'false'). Additionally there were 11,071 instances where two bots exclusively interacted (marked as 'true'). In this analysis the focus will be on the instances where genuine interactions occur between two bots.

The reason behind this choice is rooted in the objective of this study - to deeply explore the characteristics, workings, and consequences of inter-bot communication within a wider framework of crisis communication. By examining real life examples it's possible to gain insights into how social bots interact with each other to shape narratives and spread information (or misinformation) during emergency situations. Since these bots can be both harmful (by spreading information) and helpful (by countering information) understanding their patterns

of interaction is crucial. This exploration has the potential to reveal details about their strategies, actions and impact on conversations, during critical events. Consequently, it can inform the creation of more efficient strategies to handle and minimize the consequences of bot-driven false information. In order to enhance the efficiency of the analysis and maintain a strong emphasis on interactions between bots, a fresh dataset has been developed called "bot_to_bot". This specific dataset solely consists of instances that have been labeled as 'true', meaning it encompasses cases where social bots have engaged in conversations with other social bots. In the study of network properties among social bots, various characteristics both emulate and deviate from those traditionally observed in human-centric social networks. A striking feature of the network under investigation is its incomplete connectivity, signifying the existence of isolated nodes or distinct subnetworks. Such a phenomenon could point to disparate operational clusters or separate campaigns within the same overarching network. The pronounced disparity between the network's maximum degree of 492 and its average degree of 9.3 is noteworthy. This skew aligns with the 'power law' distribution often encountered in real-world social networks. Contrastingly, the network's low density and clustering coefficient of 0.27 diverge from patterns typical of human social networks. These metrics suggest a limited extent of interconnectivity and community formation, which could either serve as mechanisms to avoid detection or simply be artifacts of disjointed bot operations. Cumulatively, these observations contribute to an intricate understanding of social bot capabilities and limitations. The data indicates a nuanced approach in mimicking certain statistical attributes of genuine social networks while falling short in others. Such intricacies necessitate further research into the programming, deployment, and evolving strategies of bots to approximate human-like behaviors on social platforms.“

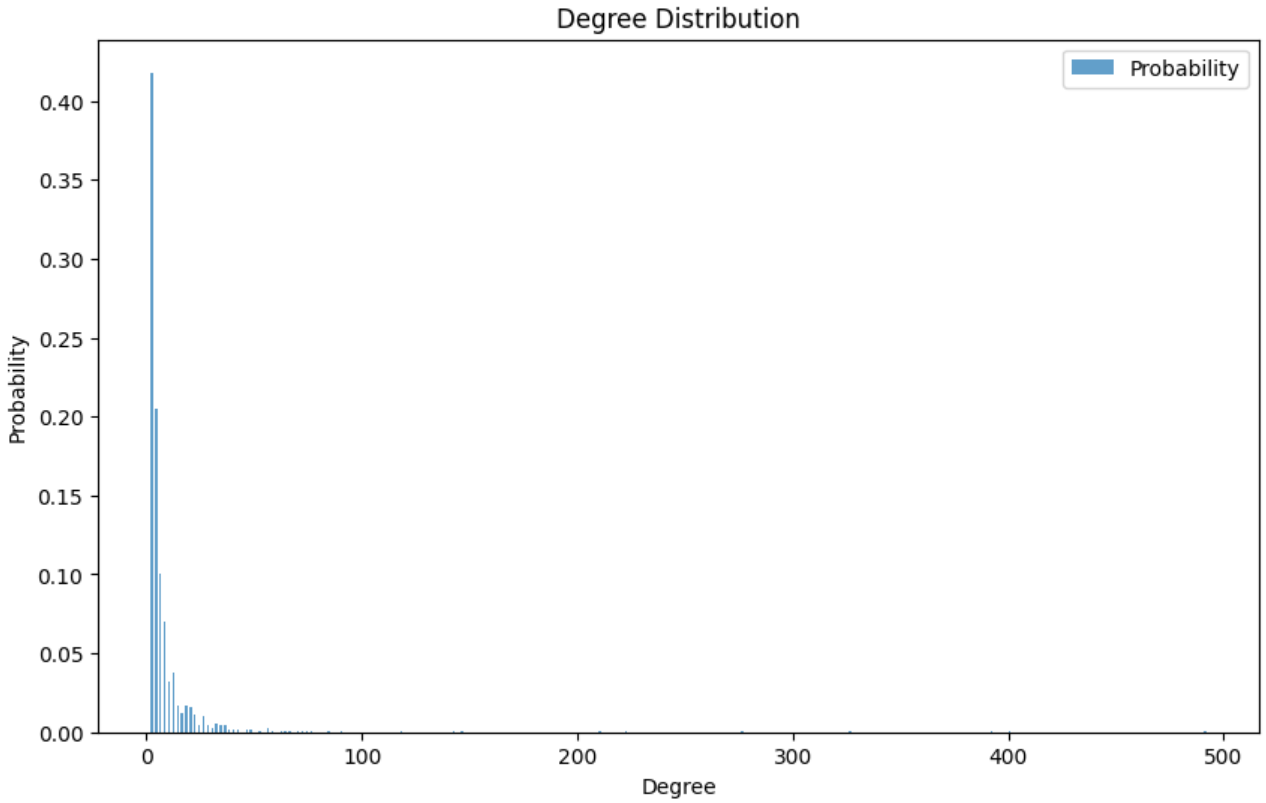

Figure 4.22: Degree Distribution

The examination of the network based on measures, like degree centrality, betweenness centrality and EigenCentrality reveals a landscape where certain nodes emerge as crucial points for the flow of information. One such node is "MATEVIDENCE," which goes beyond connectivity and plays a sociolinguistic role by bridging two distinct groups.

The significance of "MATEVIDENCE" is further emphasized by its position connecting Russian and Ukrainian speaking communities to clusters of English speaking trolls across the political spectrum. This dual language and ideological function enhances its importance as a hub for crosscultural and cross ideological interactions. Its betweenness centrality indicates its role as a pathway in the network through which a considerable amount of information must pass, granting it an unusual level of control and influence.

Disrupting or neutralizing the influence of "MATEVIDENCE" would not disrupt one of the important nodes in the network but could also potentially have a ripple effect that weakens the overall effectiveness, in spreading misinformation. This action would essentially cut off a connection, within the network hindering the exchange of information between different sub groups. The consequences of this action could be significant since the node plays a role in

allowing ideas and stories to spread and interact among separate communities.

By isolating "MATEVIDENCE " the network would lose an element that ensures not its strength but also its ability to adapt and reach a wider audience when spreading misinformation. As a result any strategies aimed at reducing the impact of this network should prioritize identifying and neutralizing targets.

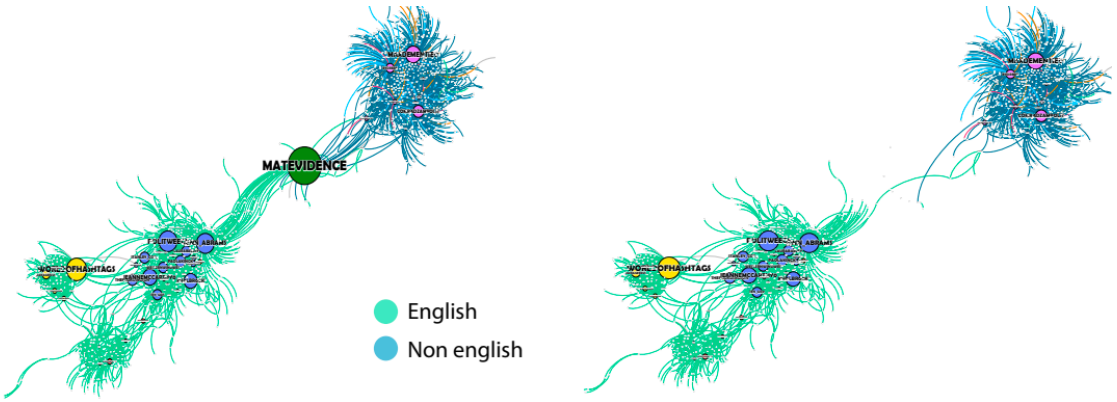

Figure 4.23: METEVIDENCE

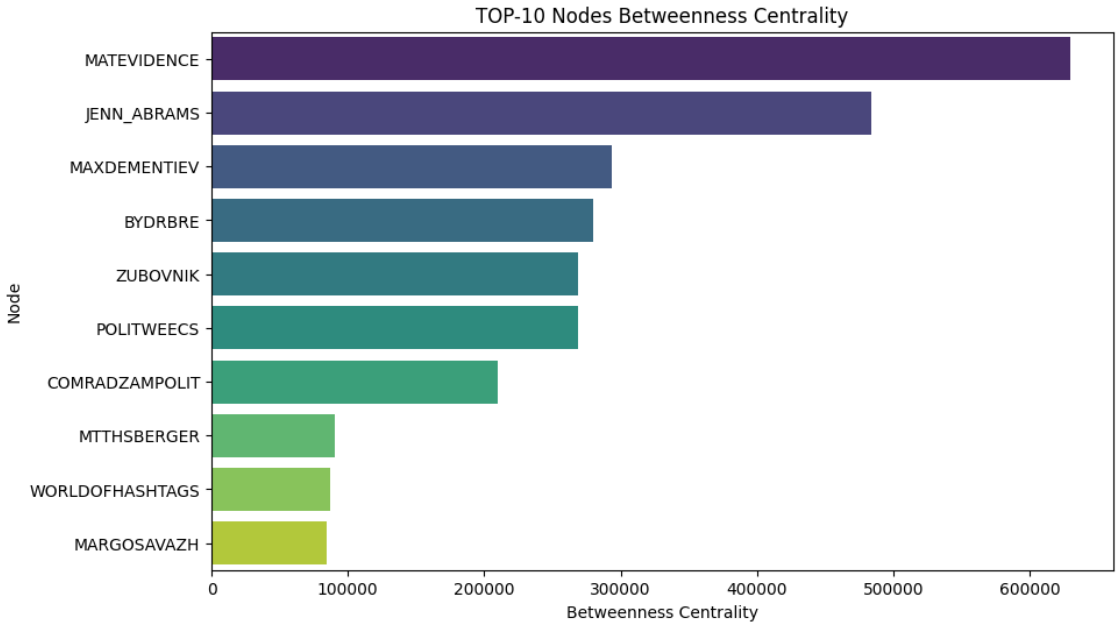

Figure 4.24: Top 10 Betweeness centrality

4.5 Key Insights and Relevance for Policy Development

In the chapter focused on identifying the problem the analysis explores three areas: content analysis, temporal analysis, and network analysis. Each of these areas provides insights that contribute to the understanding of disinformation campaigns and the approaches to combat them.

Content Analysis

Amplification Metrics and Linguistic Inconsistencies: The study's findings, on the use of "Amplification Metrics" strategies, such as counting URLs utilizing hashtags and mentioning Twitter handles provide insights into the organized methods employed to spread messages widely. This understanding allows policymakers to develop algorithms that can detect and counteract these amplification tactics thus preventing manipulation of social media platforms. Additionally the identification of "Linguistic Inconsistencies" that specifically target non native English speakers opens up opportunities, for natural language processing tools to identify anomalies. This not aids in detecting disinformation. Also contributes to the improvement of language based communication technology.

Account Category Tactics: Understanding the types of troll accounts can help policymakers better respond to disinformation. By tailoring their strategies based on the category they can effectively counter the tactics employed by these accounts. For example addressing the tactics of "RightTroll Accounts" requires an approach compared to dealing with "HashtagGamer Troll Accounts." This level of detail improves the effectiveness of policy interventions.

Temporal Analysis

Analyzing the patterns of engagement over time reveals how digital conversations change dynamically. By finding connections, between real world events and the corresponding increase, in activity it was possible to gain a strategic edge. Policymakers can use this knowledge to predict and address periods when false information spreads rapidly by implementing targeted measures to counteract it.

Network Analysis

Bot Interactions and Influence: The study's exploration of bot interactions uncovers intricate patterns in how these entities collaborate during crises. This observation underscores the need for collaborative strategies between institutions and social media platforms. Additionally, the identification of network characteristics such as incomplete connectivity and low clustering coefficients reveals the presence of operational clusters. Policymakers can leverage this understanding to craft strategies that target specific nodes, effectively disrupting disinformation dissemination pathways.

Influential Nodes: The recognition of influential nodes, exemplified by "MATEVIDENCE" illuminates the role of specific accounts in bridging diverse communities. Understanding such nodes' cross-cultural and cross-ideological influence enables policymakers to engage in targeted interventions. By neutralizing or countering the messaging from these nodes, institutions can curtail the reach of disinformation and mitigate its impact.

The understanding of tactics like "Amplification Metrics" and "Linguistic Inconsistencies" forms a cornerstone for proactive prebunking strategies. Institutions can develop educational campaigns that teach users to recognize these tactics, empowering them to critically evaluate information before sharing it. Early intervention through prebunking safeguards against the rapid dissemination of disinformation. On the other hand, insights garnered from bot interactions and the identification of influential nodes offer valuable guidance for post-incident strategies. Policymakers can devise rapid response plans to neutralize the impact of influential nodes, thereby curbing the spread of disinformation after crisis events. The ability to target specific nodes reduces the overall reach of false narratives and reinforces public trust in accurate information sources.

Incorporating these findings into policy development ensures a comprehensive, data-driven, and adaptable approach to combating disinformation's adverse effects

Chapter 5

Discussion and Conclusions

5.1 Findings

The comprehensive findings delineated in this study emerge from a meticulous analysis of empirical data, integrated coherently with a review of existing literature. This rigorous methodological approach lends credence to the outcomes, reinforcing their applicability in both academic discourse and practical interventions. The study effectively amalgamates a variety of thematic areas, weaving them into a cohesive narrative that adds depth and nuance to our understanding of disinformation in the digital age.

In elaborating upon the significance of amplification metrics and linguistic inconsistencies, the present study augments existing research frameworks concerning effective communication in the digital age. Specifically, it extends the work of Graham and Avery (2013) and Lachlan et al. (2016), who emphasized the primacy of clear and coherent messaging, by introducing the critical dimension of strategic amplification. This added layer offers institutions nuanced guidance for optimizing the dissemination of key messages, thereby improving their reach and impact. Furthermore, the study synergistically interfaces with existing literature on natural language processing, notably van der Linden et al. (2020), by showcasing how anomalies in language usage can be algorithmically detected to enhance the efficacy of prebunking strategies. In doing so, the research not only contributes a valuable tactical angle to the prevailing discourse on combating disinformation but also provides actionable insights for the development of more sophisticated, automated tools in misinformation management.

The classification and subsequent examination of types of troll accounts greatly con-

tributes to the existing body of knowledge on the participants in the dissemination of false information. Previous researchers like Conway et al. (2017) and Edwards et al. (2018) have discussed the involvement of actors. There has been a noticeable gap when it comes to understanding the nuanced tactics employed by these specific troll accounts. This study fills that gap by presenting a classification system for troll accounts including categories such as 'RightTroll' and 'HashtagGamer,' and outlining strategies to counter each category.

For example 'RightTroll' accounts typically engage in polarizing discussions focusing on promoting narratives driven by ideology. As a result appropriate counter strategies for this category might involve targeted use of fact checking resources or employing sentiment analysis algorithms to identify and scrutinize polarizing messages closely. Conversely 'HashtagGamer' accounts are often less concerned, with ideology. Instead concentrate on manipulating social media algorithms to amplify their disinformation campaigns. Strategies to combat this category could be algorithmically oriented, focusing on adjusting platform dynamics to prevent content boosting.

Interventions targeted at categories significantly contribute to the growing field of tactics to combat disinformation. These interventions offer insights and recommendations that go beyond suggestions benefiting policymakers and researchers. It becomes evident that countering disinformation requires accounting for the objectives, mechanisms and tactics used by troll accounts.

By adding this level of detail the study enhances the discussions about fighting against disinformation. Furthermore these insights provide a toolkit for those involved in formulating policies to counter disinformation whether they are government bodies, social media companies or civil society organizations. While previous academic discussions have hinted at the need for a nuanced approach towards actors involved in disinformation this study not aligns with those discussions but also advances them by offering concrete strategies to address the diverse landscape of disinformation actors. This represents a step towards understanding, ultimately reducing the impact of the disinformation epidemic.

The analysis of time based connections, between real world events and the rise of disinformation, enhances the urgency previously highlighted in research conducted by Woolley and Howard (2016) and Tucker et al. (2018). While these scholars have shed light on the spread of disinformation during critical moments there remains a gap in understanding the

temporal mechanics that can enable proactive countermeasures. This study significantly advances our knowledge by providing an understanding of the timing factors serving as a tool for policymakers.

This connection holds value beyond academia. Policymakers can leverage these insights to anticipate periods when disinformation's likely to surge such as during elections, natural disasters or other high stakes events. By deploying counter resources this study essentially offers an "early warning system" based on observed correlations, between specific real life events and increased disinformation activity.

Moreover this predictive tool has the potential to revolutionize reactive strategies employed against disinformation. Of playing catch up and attempting damage control after its spread policymakers can now take proactive measures.For instance one approach could be to plan information campaigns or increase surveillance during anticipated periods of high disinformation activity. This shift, from a reactive to a stance could have an impact on combating disinformation. It would enhance the effectiveness of countermeasures, potentially discouraging those who spread disinformation from initiating campaigns in the place.

Furthermore this study introduces a solution that fills a gap in research. While concerns about the spread of disinformation have been well documented by researchers like Vosoughi et al. (2018), there has been limited exploration of measures to address these concerns. By translating analysis into strategies this study not only confirms the urgency emphasized in existing literature but also equips policymakers with practical tools to act proactively and effectively.

In the realm of network analysis, the present study's examination of bot interactions and influential nodes serves as a significant addition to existing literature. While scholars like Ferrara et al. (2016) and Shao et al. (2018) have scrutinized the role of automated systems in the diffusion of disinformation, the current work extends the discourse by incorporating a much-needed policy perspective. Specifically, the study posits that tackling the issue of bot-generated disinformation necessitates a collaborative approach between governing institutions and social media platforms. This articulation of the policy implications fills a vital gap left by prior studies, such as Bodo et al. (2019), that did not emphasize actionable policy measures for mitigating the influence of bots.

In the same vein, the work sheds light on another crucial but underexplored component: influential nodes within social networks. While network behavior has been analyzed by Conway et al. (2017) and Edwards et al. (2018), those analyses often overlooked the importance of identifying and targeting specific, influential nodes in the network. These nodes, as the study demonstrates, serve as bridges between disparate communities and therefore wield disproportionate influence in the spread of disinformation. By focusing on these key nodes, the current research offers a targeted approach for intervention, a strategy that aligns with crisis communication principles elucidated by Reynolds and Seeger (2005) and Lachlan et al. (2016). In essence, targeting influential nodes could serve as an effective mechanism for restoring public trust and mitigating the impact of disinformation campaigns.

This nuanced understanding is pivotal, not only for academia but also for policy formulation and implementation. By examining both bot interactions and influential nodes, the research provides a multi-faceted approach for combating disinformation. Policymakers can leverage these insights to formulate robust, data-driven strategies that account for both automated and human-centric elements in the disinformation ecosystem. Whether it's by implementing regulations on social media platforms to control bot activity or by developing public campaigns that target influential nodes, the study provides a roadmap for multifaceted interventions.

5.2 Establishing Solution Goals and Objectives

The growing impact of bots in communication poses a challenging and multifaceted problem that requires a strong evidence based approach for effective management. The core of the matter lies in the fact that bots have the ability to either amplify messages or distort them through disinformation tactics. It is crucial to comprehend the functionalities and limitations of bots in order to develop strategies that can minimize their harmful consequences while utilizing their capabilities for beneficial purposes.

As such, it has become imperative for organizations to implement an effective policy for managing the risk and impact of disinformation. When it comes to being proactive taking measures to counter disinformation, also known as prebunking, can make an impact. Educating and raising awareness about bots tactics can help people become more discerning when encountering messages. Additionally, using advanced machine learning algorithms to monitor communication channels in time can help identify bot activity. These algorithms

have improved in their ability to distinguish between malicious bots allowing for precise management strategies.
When it comes to dealing with disinformation campaigns effective communication after an incident is also of importance. Institutions need to act transparently providing evidence based information to regain trust. Attribution techniques play a role in this process by using algorithms to track the source of disinformation.
Managing the influence of social bots in institutional communication necessitates a meticulously crafted, dynamic policy framework that blends proactive and reactive strategies, supported by cutting-edge technological solutions and continual evaluation mechanisms.
The aim of the policy is to empower institutions to proactively and reactively manage the risk and impact of disinformation propagated by social bots, thereby safeguarding institutional credibility and the integrity of the information ecosystem.

5.2.1 Structure of the Policy

The policy guideline for institutions, aims at enhancing the management of social bots and disinformation. This policy will incorporate both prebunking strategies and post-incident communication.

- **Policy Objective**: To proactively and reactively manage the risk and impact of disinformation propagated by social bots.
- **Prebunking Strategy**: Educational programs, awareness campaigns, and periodic briefings on new tactics of social bots and disinformation strategies.
- **Monitoring and Detection Guidelines**: Continuous monitoring of institutional communication channels to identify disinformation signals and the use of AI-based tools to detect bot activity.
- **Post-Incident Communication Plan**: Actions that an institution should undertake following a disinformation incident and the importance of timely, transparent, and clear communication
- **Post-Incident Evaluation**: Review and learning from each incident and strategies to evaluate the efficacy of the post-incident response and prebunking strategy.

- **Regular Updates and Revisions**: The policy should be consistently updated based on new insights, technologies, and tactics.
- **Policy Enforcement and Compliance**: How the policy will be applied and the implications for non-compliance.

5.2.2 Structure and Distribution of Responsibilities

Addressing such a complex and pervasive issue requires a holistic, multidisciplinary, and well-coordinated approach. However, in the interest of simplicity and greater efficiency, the responsibilities can be streamlined into four key sectors within organizations:

Policy Objective

Executive Leadership: The objective is to proactively and reactively manage the risk and impact of disinformation propagated by social bots. The leadership tier could allocate specific budgets and resources to implement this policy and ensure that it aligns with the company's broader risk management objectives.

Prebunking Strategy

Communications/Public Relations: This strategy involves the creation of educational programs and awareness campaigns. Periodic briefings can be scheduled to update staff on emerging tactics used by social bots and disinformation strategies. This department should work in close collaboration with the IT/Cybersecurity officers to ensure that educational content is up-to-date and effective.

Monitoring and Detection Guidelines

IT/Cybersecurity Officers: These guidelines could involve implementing AI-based tools that continuously monitor institutional communication channels to identify any disinformation signals. Officers can also use machine learning algorithms to detect anomalous bot activity and flag it for review.

Post-Incident Communication Plan

Communications/Public Relations: After an incident, this department should have a set of predefined actions to ensure that information is disseminated transparently and clearly.

This could include issuing official statements, FAQs, and briefings to both internal stakeholders and the public to mitigate the impact.

Post-Incident Evaluation

Legal Department: Following each incident, there should be a thorough review and learning process that involves a legal analysis. This will help in assessing the efficacy of the response and determining if any legal liabilities have arisen due to the incident.

Regular Updates and Revisions

Executive Leadership: Regular reviews should be scheduled to update the policy. These reviews can incorporate new insights, technologies, and tactics that have been learned either through internal experiences or from external developments in the field.

Policy Enforcement and Compliance

IT/Cybersecurity Officers and Legal Department: Both departments should work together to set up technical controls that enforce the policy and ensure compliance. They should also clarify the implications of non-compliance, which could range from internal disciplinary action to legal consequences.

5.3 Prebunking Strategy

In the increasingly digital landscape, communication departments within various organizations face the escalating issue of disinformation. Their role is pivotal in molding both public perception and internal understanding of various subjects, making them gatekeepers of credibility and trust. With disinformation tactics growing more sophisticated through advanced social bots and algorithmic content distribution, a multi-layered defensive strategy is indispensable. A key element in this multi-pronged approach is the utilization of prebunking — a proactive measure that aims to educate staff and stakeholders to recognize and counteract misinformation before it takes root. The strategy proposed here aims to be a defense by equipping institutions with a comprehensive set of tools, skills and knowledge in a proactive manner. This multi faceted approach combines targeted initiatives, with hands on exercises to create a well rounded framework that goes beyond traditional counter strategies.

Responsibility Area	Dept.	Tasks
Policy Objective	Exec. Leadership	Allocate budgets, align policy
Prebunking	Comm./PR	Create programs, brief on tactics
Monitoring	IT/Security	Implement AI tools, monitor channels
Post-Incident Comm.	Comm./PR	Issue statements, FAQs
Post-Incident Eval.	Legal	Conduct legal analysis
Updates	Exec. Leadership	Schedule reviews
Compliance	IT/Security & Legal	Set controls, clarify implications

Table 5.1: Division of Responsibilities for Managing Social Bots and Disinformation

Real life examples are included to give context providing insights into disinformation campaigns that have been successful. These real world case studies act as reference points allowing for a grasp and long term retention of important concepts. The educational modules are based on principles drawn from science. The aim is to improve analytical skills strengthening individuals and communities against disinformation tactics.
Community involvement is given importance in this strategy. It recognizes that tackling disinformation is not an individual problem but a collective one. Therefore it promotes the sharing of knowledge within communities, encouraging efforts to address this complex issue. The comprehensive and multi level prebunking approach outlined in this strategy enhances its adaptability and effectiveness making it an invaluable tool in the fight against the forms of disinformation.

Based on research findings regarding "Amplification Metrics" and "Linguistic Inconsistencies " this section provides guidance for implementing the research insights. It offers a step by step plan for developing initiatives and online campaigns that aim to train individuals in recognizing disinformation tactics. The blueprint recommends using real life examples and online tutorials to illustrate how social bots use hashtags and URLs to amplify messages serving as a heuristic for identifying disinformation campaigns. Moreover it

takes into account the study's identification of inconsistencies which can be misleading for native English speakers. In light of this the proposal suggests training modules grounded in science principles to enhance analytical skills among these vulnerable groups. Together, these multi-layered, evidence-based educational programs not only enhance individual-level critical thinking but also build collective resilience against disinformation. This nuanced strategy addresses both the dissemination mechanisms of false narratives and the cognitive vulnerabilities they exploit, making it a comprehensive and adaptable tool in the evolving landscape of disinformation.

The process of prebunking to counteract disinformation necessitates a meticulous, multi-layered approach to be truly effective. The aim is to arm institutions with the skills and knowledge to discern false information before it proliferates.

5.3.1 Amplification Metrics

The "Amplification Metrics" section outlines a multi-faceted approach aimed at enhancing organizations' abilities to understand and counter disinformation campaigns that exploit social media algorithms.
The methodology is constructed over four principal stages: Educational Programs, Real-World Case Studies, Interactive Exercises, and Certification.

Firstly, the Educational Programs serve as the cornerstone for disseminating foundational knowledge. These should be incorporated into existing digital literacy curricula or made part of induction programs. The curriculum must include an in-depth understanding of how social media algorithms function, with particular focus on how they can be exploited through tactics such as 'Amplification Metrics.' Given the highly specialized nature of this subject, reference to seminal studies on social bot tactics could augment the educational rigor.

Secondly, Real-World Case Studies are integrated into the curriculum to apply theoretical knowledge to practical scenarios. The case studies should be carefully selected to illustrate instances where 'Amplification Metrics' and 'Linguistic Inconsistencies' were evidently manipulated to propagate disinformation. By scrutinizing these cases, participants can more deeply understand how to recognize these tactics in a live environment. This method of experiential learning, backed by empirical research, will add a layer of credibility and urgency to the course.

Thirdly, Interactive Exercises must be developed to reinforce theoretical learning with hands-on experience. These could be implemented in the form of digital simulations where users are tasked to identify posts or content that utilize these manipulative strategies. Given that we are dealing with an expert audience, these exercises should not only test recognition skills but also the analytical capacity to determine the probable impact of such tactics on information dissemination.

Finally, a Certification process will be put in place to evaluate the mastery level of participants. This certification, ideally endorsed by a reputable institution, would lend weight to the program and provide tangible evidence of expertise in countering disinformation. Moreover, this final step could serve as a pre-requisite for stakeholders to engage in higher-level counter-disinformation activities, thereby ensuring that only the most capable are entrusted with this critical responsibility.

Each of these stages contributes to a robust, evidence-based prebunking strategy aimed at effectively mitigating the impacts of disinformation campaigns.

Educational Programs

The main objective of this program should be to explain how social media algorithms work and their weaknesses. This is crucial because these algorithms are often the foundation for spreading disinformation campaigns.
It is crucial to understand the role that these metrics play in the world of content sharing and visibility. Likes, shares, mentions, URL sharing and hashtag usage are not just indicators of content popularity but tools that can be manipulated to distort conversations.
First and foremost combining URLs with hashtags is a strategy used to quickly spread messages to a wide audience. The course should delve into the details of how certain URLs,especially shortened ones, can mask the origins of disinformation. Additionally participants should learn how tracking the frequency and spread of these URLs can provide insights into organized campaigns.
Moreover the use of hashtags is an aspect. While hashtags were initially created as a way to categorize content around topics, they are now exploited to hijack trending discussions and create artificial trends. Participants need to grasp how hashtags propagate and how malicious actors can manipulate them to lead users towards disinformation.

The course must also highlight the significance of mentions. Demonstrate how tagging or mentioning users or organizations helps amplify content. This technique aims to leverage their reach in order to lend credibility to disinformation campaigns. It is equally important to explore how these metrics are interconnected. For example a post that has an URL might receive a lot of likes quickly because of hashtag promotion. This in turn prompts the algorithm to consider it as relevant content, which leads to more promotion. By looking at it this way participants will understand how different metrics can work together for impact.
By examining the complexities of amplification metrics in detail stakeholders will gain the knowledge needed to recognize the strategies used by disinformation campaigns and take steps to counteract them. This will create a foundation, for modules enabling them to apply and test their newfound insights effectively.

Real-world case scenarios

When it comes to the "Real World Case Studies" it's possible to delve into how these cases help participants grasp the concept of amplification metrics on a level. By incorporating real world case studies into programs examples that reinforce theoretical principles are provided and deepen understanding of how these metrics are manipulated to spread disinformation.
In the training module case studies from incidents where hashtags, URLs or mentions were artificially inflated to amplify a message are worthy. For example lets explore a disinformation campaign during an election and uncover how non organic tweets used trending hashtags to gain visibility. It's possible to lay out metrics,such as tweet velocity, time stamps and the history of user accounts for participants to analyze closely.
Participants will be guided in identifying anomalies like retweet rates, frequent occurrences of the same URL or suspicious bursts of activity around particular hashtags. To make it more realistic and complex these real world case studies can also include supporting materials, like reports or academic articles that validate the anomalies found in these amplification metrics.
The objective of this program segment is to educate stakeholders on how to differentiate between social media engagement and intentionally orchestrated strategies driven by automated bots. Real life instances serve as resources for strengthening this skill set as they require participants to apply knowledge in analyzing real cases of disinformation campaigns. This hands on approach boosts both memory retention and practical application equipping participants, with the tools to identify and combat disinformation, in real time situations.

Interactive Exercises

The "Interactive Exercises" component is crucial for practical application and retention of the concepts discussed in the educational program. Interactive online simulations can offer participants a safe space to apply their newly acquired skills on amplification metrics, without the risk of disseminating disinformation further.
For instance, one exercise could be a simulated Twitter dashboard where participants are tasked with identifying tweets that display telltale signs of bot-driven amplification. The simulation could feed a mix of genuine and bot-generated tweets, with varying degrees of subtlety in their amplification tactics. Metrics like the speed of retweets, the use of trending but irrelevant hashtags, and irregular time-stamp patterns could be programmed into the simulation.
Another exercise could be a "spot the difference" game where participants compare two similar posts side-by-side to identify which one is using amplification metrics for disinformation. This could be particularly useful for understanding linguistic inconsistencies and how they can also be a form of amplification.
To make it even more impactful, a real-time scoring system could be integrated, providing instant feedback on the participant's performance. This would not only gamify the experience but also offer a measure of one's proficiency in identifying these tactics.
Upon successful completion, the system could provide a breakdown of the correct and incorrect choices made by the participant, along with explanations. This 'review' feature serves as another learning opportunity, reinforcing the right methods for identifying amplified content while correcting misunderstandings or gaps in knowledge. The goal here is to transition participants from theoretical understanding to practical application, ensuring that they are not just aware of how amplification metrics work, but are also capable of identifying them in a real-world context. By combining educational content, real-world case studies, and interactive exercises, the program aims to provide a holistic, hands-on learning experience that effectively equips stakeholders to combat disinformation.

Certification

The "Certification" component serves as a way to evaluate participants skills and provide evidence of their competency. It acts as a culmination of the training program offering an assessment process that adds credibility and encourages adoption among stakeholders.
To create a certification process there can be assessment formats utilized. These may include multiple choice questions, scenario based assessments and time sensitive simulations

resembling the exercises covered earlier in the program. For instance participants could be presented with a collection of social media posts. Asked to identify which ones utilize amplification metrics within a given timeframe. This mirrors the urgency often required in countering disinformation campaigns.

To ensure the certification holds value it is advisable to collaborate with institutions or professional organizations that can accredit the program. This external validation adds weight to the certificate making it an impactful addition, to ones portfolio.

Additionally it is beneficial to implement levels" of certification corresponding to varying skill sets or depths of knowledge. Level 1 may cover identification skills while higher levels delve into advanced analytics and counter strategies. It provides participants with opportunities, for continual learning and specialization.

Lastly, introducing a system, for renewing the certification could be an approach to ensure that stakeholders stay up to date with their skills. As strategies for spreading misinformation and amplifying it continue to evolve it is crucial for the certification program to adapt accordingly. This can be achieved by requiring participants to undergo recertification every years in order to maintain their standing.

By establishing an accredited certification process the program goes beyond education; it sets a quantifiable benchmark against which stakeholders can assess their skills. This offers them a means of contributing to the ongoing battle, against disinformation.

5.3.2 Linguistic Inconsistencies

The "Linguistic Inconsistencies" part of the prebunking strategy focuses on addressing the ways language is manipulated in disinformation campaigns. The aim is to create training modules to improve the reading skills of those involved with a specific emphasis, on understanding linguistic nuances. This kind of training is especially useful for individuals who are not English speakers and may miss these subtle cues that indicate misinformation.

By incorporating Natural Language Processing (NLP) tools into these modules it's possible to automate the process of detecting language manipulations making it more efficient and accurate. Providing NLP tools, as browser extensions or mobile apps would allow users to easily utilize them in time while browsing the internet. For example users could receive alerts when they encounter phrases or sentence structures commonly associated with disinformation.

To further enhance the effectiveness of these modules it's noteworthy integrate principles from science. By designing gamified approaches based on cognitive science research the learning

process becomes more interactive and engaging. For instance a game could present users with statements. Ask them to identify which ones contain linguistic inconsistencies.
Moreover developing a community platform where stakeholders can share examples of inconsistencies they have come across would be beneficial well. By gathering information and debunking together as a community this platform has the potential to be a resource, for learning and staying updated on the tactics used in disinformation campaigns. This aspect of the prebunking strategy, which focuses on addressing language inconsistencies serves as a training tool, for individuals involved. It helps them recognize, comprehend and effectively counteract this element of disinformation campaigns.

5.4 Monitoring and Detection Guidelines

Grounded in the results of an extensive network analysis, it becomes imperative for organizations to recognize the role of nodes with high betweenness centrality in the dissemination of disinformation. These nodes act as pivotal 'bridges' that connect otherwise disparate segments of the network, enabling the accelerated spread of false narratives. The identification of these nodes offers actionable intelligence that can be used to mitigate the spread of disinformation effectively.

For the IT/Cybersecurity department, this means instituting a robust set of protocols to constantly monitor these key nodes. Leveraging network analysis findings, advanced analytical methods, devoid of AI-based solutions, can be employed to scrutinize activities on these nodes. Through manual scrutiny and real-time monitoring protocols, anomalies in network behavior can be promptly identified.

However, monitoring is only the first step. The temporal trends highlighted through rigorous temporal analysis should also be integrated into the monitoring process. Special attention should be given during time frames correlated with a high likelihood of disinformation campaigns, often influenced by real-world events. By incorporating such data-driven insights into monitoring protocols, the IT/Cybersecurity department can optimize resource allocation and apply targeted scrutiny when it is most needed.

Once potential disinformation-spreading nodes are identified, the next critical step involves inter-departmental coordination, primarily with the Communications/Public Relations

department. This is essential for a two-fold reason: Firstly, to engage in preemptive 'pre-bunking' measures, aimed at educating both internal stakeholders and the public, thereby making them more resilient to disinformation tactics. Secondly, to ensure that identified nodes are promptly reported to the social media platforms where they exert influence.

This report-and-act strategy allows the Communications/Public Relations department to prepare and activate their post-incident communication plans, which may include issuing official statements and FAQs aimed at countering the disinformation. It's not merely about crisis management; it's about proactive engagement and targeted intervention to minimize the impact and reach of disinformation.

Through a multidisciplinary approach that involves seamless coordination between IT/Cybersecurity and Communications/Public Relations departments, organizations can substantially enhance their ability to combat disinformation. All of these actions should be executed as part of a broader, organization-wide policy that aligns with the most recent findings from network and temporal analyses

5.5 Post-Incident Communication Plan

In the wake of a disinformation incident, the Communications/Public Relations department carries a pivotal role in restoring trust and clarity, both internally and externally. This department is equipped with a predefined set of actions designed to ensure transparent and clear dissemination of information. Such actions can include the immediate release of official statements that refute the disinformation, along with Frequently Asked Questions (FAQs) to address common queries and concerns. These materials are prepared in collaboration with IT/Cybersecurity and Legal departments to ensure accuracy and compliance. Internally, briefings may be conducted to educate stakeholders on the incident's nature and the steps being taken to resolve it. Externally, these communications aim to quell any burgeoning narratives fueled by the disinformation, effectively neutralizing its impact. Moreover, the Communications/Public Relations department is responsible for liaising with social media platforms to halt the spread of disinformation at its source, especially concerning nodes identified through network analysis as being high-risk vectors for the spread of false information. Through a well-executed Post-Incident Communication Plan, the organization not only mitigates the immediate impact but also lays the groundwork for long-term strategies to combat

future disinformation campaigns.

5.6 Post-Incident Evaluation

The Legal Department plays a crucial role in the aftermath of a misinformation incident by performing a thorough post-incident evaluation. As these divisions offer a thorough method of evaluating the issue, this evaluation is organized to replicate the original research framework, divided into content analysis, temporal analysis, and network analysis. The Legal Department works closely with the IT/Cybersecurity, Communications, and Public Relations departments to comprehend the entire breadth and effects of the occurrence. A particular emphasis is placed on making sure that the actions implemented, both in the mitigation process and the subsequent public communication, are in line with all applicable rules and regulations. The department may examine matters such as data protection violations and libel as well as other legal hazards that could put the company at risk. The findings of this comprehensive evaluation serve two purposes: they inform changes to current policies, strengthening their defenses against subsequent attempts at disinformation, and they offer actionable legal insights that may be important in pursuing legal actions against the campaign's perpetrators. The Legal Department plays a crucial part in the continual development and reinforcement of the organization's overall misinformation counter-strategy by using a disciplined, research-based approach to each post-incident investigation.

5.7 Regular Updates and Revisions

The Executive Leadership bears the ultimate responsibility for ensuring that the organization's disinformation countermeasures remain effective, current, and responsive to an ever-evolving landscape. Regular reviews are therefore imperative and should be scheduled at least semi-annually, if not more frequently, depending on the scale and speed of external developments. These reviews aim to reassess the policy in light of new insights, technological advancements, and tactical shifts that may have occurred both within the organization and in the broader context of disinformation warfare.

To facilitate this, executive leadership should engage with IT/Cybersecurity Officers, Communications/Public Relations, and the Legal Department to form a multidisciplinary review team. This team is tasked with critically examining the existing strategy and identi-

fying gaps or vulnerabilities that have either been exposed during actual incidents or could be predicted based on the latest research and case studies.

Key Performance Indicators (KPIs) should be an integral part of these reviews. Metrics such as the rate of detected disinformation incidents, the effectiveness of prebunking strategies, or the timeliness and impact of post-incident communications could serve as quantifiable measures. These KPIs should be benchmarked against industry standards and assessed in the context of real-world results.

Moreover, given the dynamic nature of disinformation tactics, the review team must stay abreast of advancements in AI and machine learning for detection and prevention. New technologies could offer more robust solutions for monitoring nodes with high betweenness centrality, a critical factor identified in previous network analyses.

The outcome of these reviews should serve dual purposes. Firstly, immediate revisions may be required to address any current vulnerabilities in the existing policy. Secondly, the insights should feed into a strategic roadmap designed to fortify the organization's long-term resilience against disinformation. In summary, these reviews are not one-off exercises but cyclical processes that serve as stepping stones for continuous improvement, adaptability, and resilience in the fight against disinformation.

5.8 Policy Enforcement and Compliance

Both the IT/Cybersecurity Officers and the Legal Department have critical roles to play in the enforcement of the organization's disinformation policy and ensuring its compliance. These departments should work synergistically to create a robust framework that integrates both technical and legal controls to effectively mitigate the risks associated with disinformation.

On the technical front, IT/Cybersecurity Officers are responsible for implementing safeguards, such as real-time monitoring systems, AI-based detection algorithms, and other security measures, to prevent the dissemination of disinformation. These technical controls should be designed to actively identify nodes with high betweenness centrality, as these are often critical in the spread of false narratives, a conclusion reached based on previous network

analysis findings.

On the legal side, the Legal Department needs to provide comprehensive guidelines on how the organization should respond when policy breaches occur. This may include the formulation of internal disciplinary procedures and actions, escalation matrices, and legal consequences that could be enforced in case of serious offenses.

The implications for non-compliance should be well-documented and disseminated across the organization to ensure that all employees are aware of their roles and responsibilities. Clear consequences for non-compliance could range from internal disciplinary action such as warnings, suspensions, or reassignments, to external legal consequences that could involve litigation or other penalties.

Regular audits and assessments should be conducted to ensure that the implemented technical controls are in compliance with the policy and that they are effective in meeting the organization's objectives in combating disinformation. Any gaps identified during these audits should be jointly discussed by IT and Legal departments, and corrective measures should be implemented promptly.

In addition, the IT/Cybersecurity and Legal teams should meet periodically to review the effectiveness of the policy enforcement mechanisms in place and recommend updates based on new legal developments, technological advances, and any lessons learned from incidents of disinformation affecting the organization.

In essence, policy enforcement and compliance is a collaborative, ongoing effort that requires the specialized skills and expertise of both the IT/Cybersecurity Officers and the Legal Department. By working together, these departments can provide a more holistic, effective approach to combating disinformation within the organization.

5.9 Inter-Departmental Synergy in Combatting Disinformation: Roles and Coordination

The comprehensive policy to combat misinformation includes a framework that clearly defines duties and responsibilities for each department. The Executive Leadership maintains congru-

ence with company goals and is responsible for overseeing the overall policy objective. Prebunking plans and post-incident communications are led by communications and public relations, keeping both internal stakeholders and the general public informed. IT/Cybersecurity Officers concentrate on real-time monitoring and misinformation activity identification, particularly when it involves high-risk network nodes. The Legal Department evaluates incidents afterward to determine any potential legal repercussions. The symbiotic link between the IT/Cybersecurity and Communications/Public Relations departments, which is neatly illustrated by a bi-directional grey dashed line labeled "Coordination for Prebunking & Post-Incident Plans," is a critical component of this multi-tiered approach. The ability of the company to proactively and reactively control the risks posed by disinformation campaigns is optimized by this mutual collaboration.

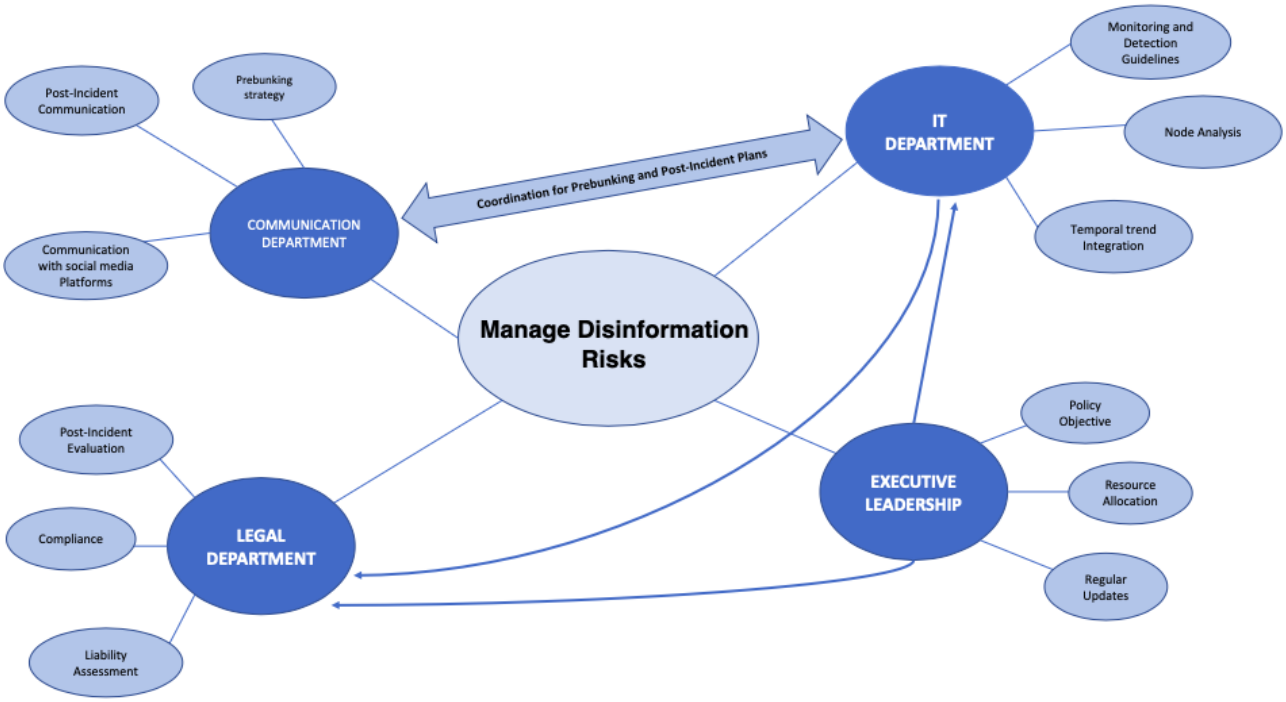

Figure 5.1: Inter-Departmental Coordination

5.10 Conclusions

The results of this study provide a relevant contribution to the existing research by presenting an approach to combatting disinformation. Unlike studies that focused on reactive or proactive measures this study advocates for a balanced and flexible strategy that incorporates both. By addressing the gap in the literature, which lacks a unified approach to tackle

disinformation this research offers solutions for policymakers, strategists and social media platforms.

What makes this research unique is its emphasis on actionability.The research goes beyond discussing theories. Offers practical insights that various individuals and organizations involved in sharing information and creating policies can use right away. The significance of this is extremely important, in todays time when disinformation campaigns are becoming more complex and influential.

In addressing the research questions, the findings offer several noteworthy implications for the academic and professional communities.

Data shows that AI-based monitoring systems are excellent at spotting abnormal activities that may be attributed to social bots, which is important for the effective management of post-incident disinformation. This implies that the incorporation of such systems into institutional communication infrastructures is not only advantageous but also necessary for the prompt thwarting of disinformation efforts. The data also indicates that if these systems aren't updated to account for changing bot techniques, their effectiveness would decline. This emphasizes how important it is to keep making technology investments.

Moreover, the study finds that public engagement significantly increases the effectiveness of anti-disinformation measures in institutional communication strategies. It implies that vigilant community monitoring of potential misinformation serves as a strong defense. The data also shows, meanwhile, that the lack of control on public platforms might open up new channels for misinformation, stating

The results demonstrate that disinformation operations frequently use psychological cues to maximize involvement. An emphasis on more evidence-based policy interventions is signaled by the gathering of empirical data in order to pinpoint these triggers. As a result, the evidence supports the use of a multidisciplinary strategy that combines behavioral psychology and data science to create effective policies.

These discoveries have a wide range of consequences. They make the case for a comprehensive, data-driven policy framework in addition to the implementation of certain tactical improvements. The ideal framework would integrate technology defenses, citizen involve-

ment, and an understanding of misinformation tactics that is based on empirical research. The conclusions thus demand for additional investigation into these areas in order to improve the methods for successfully battling misinformation.

5.11 Limitations and Scope for Future Research

While this study provides valuable insights into combating disinformation through a multi-faceted approach, it is not without its limitations. First, the focus on nodes with high betweenness centrality, although important, might not cover the full spectrum of network entities involved in the spread of disinformation. Second, although advanced analytical methods have been employed, the absence of AI-based solutions in this study could be considered a limitation, given the evolving sophistication of disinformation campaigns. The study also assumes a degree of inter-departmental coordination that may not be feasible in all organizational settings. Additionally, the scope of this research is limited by the types of disinformation campaigns examined; for instance, tactics involving Large Language Models (LLMs), were not extensively considered. These limitations open avenues for future research, which could focus on the role of emerging technologies in disinformation campaigns, as well as the effectiveness of various types of organizational structures in combating such efforts. Future work should also aim to empirically test the strategies recommended in this study across diverse settings to ascertain their generalizability and effectiveness.

Bibliography

(2018). Ethics guidelines for trustworthy ai. Technical report, European Commission.

(2023). Attacco hacker in italia, tim down.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Badawy, A., Ferrara, E., and Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, pages 519–528.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion.

Bodo, L., Helberger, N., and Irion, K. (2019). Regulating disinformation with artificial intelligence. *Information Communication Society*, 22(5):689–704.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24(2):505–528.

Committee, R. P. (2021). The solarwinds cyberattack.

Conway, B. A., Kenski, K., and Wang, D. (2017). The rise of twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary. *Journal of Computer-Mediated Communication*, 22(4):232–248.

Couldry, N. and Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3):398–415.

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., and Johnson, B. (2019). The tactics & tropes of the internet research agency. *Insert Journal Name*, Insert Volume Number(Insert Issue Number):Insert Page Numbers.

Edwards, C., Edwards, A., Spence, P. R., and Shelton, A. K. (2018). Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33:372–376.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. (2015). "i always assumed that i wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 153–162. ACM.

Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *arXiv preprint arXiv:2004.09531*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Gillespie, T. (2014). *The relevance of algorithms*, pages 167–194. MIT Press.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.

Gorwa, R. and Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2):225–248.

Graham, M. and Avery, E. J. (2013). Government public relations and social media: An analysis of the perceptions and trends of social media use at the local government level. *Public Relations Journal*, 7(4):1–21.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Howard, P. N. and Kollanyi, B. (2016). Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. Technical Report 2016.2, Project on Computational Propaganda Research Paper.

Jamieson, K. H. (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press.

Jurafsky, D. and Martin, J. H. (2019). Speech and language processing. Draft of September 23, 2019.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12.

Lachlan, K. A., Spence, P. R., Lin, X., Najarian, K., and Del Greco, M. (2016). Social media and crisis management: Cerc, search strategies, and twitter content. *Computers in Human Behavior*, 54:647–652.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lewandowsky, S., Ecker, U. K., and Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369.

Marwick, A. and Lewis, R. (2017). Media manipulation and disinformation online.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):205395171667967.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

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

Pennycook, G. and Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*.

Reuters (2023). Italy sounds alarm on large-scale computer hacking attack.

Reynolds, B. and Seeger, M. (2005). Crisis and emergency risk communication as an integrative model. *Journal of Health Communication*, 10(1):43–55.

Rogers, E. M. (2003). *Diffusion of Innovations*. Free Press, 5 edition.

Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world*. W. W. Norton & Company.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Solove, D. J. (2008). *Understanding Privacy*. Harvard University Press.

Starbird, K. (2019). Disinformation's spread: Bots, trolls and all of us. *Nature*, 571(7766):449–454.

Starbird, K., Spiro, E., Edwards, I., Zhou, K., Maddock, J., and Narasimhan, S. (2018). Could this be true? i think so! expressed uncertainty in online rumoring. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political Science*.

Tufekci, Z. (2015). Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *Colorado Technology Law Journal*, 13:203.

van der Linden, S., Roozenbeek, J., and Compton, J. (2020). Inoculating against fake news. *Social Psychological and Personality Science*.

Veil, S. R., Buehner, T., and Palenchar, M. J. (2011). A work-in-process literature review: Incorporating social media in risk and crisis communication. *Journal of Contingencies and Crisis Management*, 19(2):110–122.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking.

Woolley, S. C. and Guilbeault, D. (2017). Computational propaganda in the united states of america: Manufacturing consensus online.

Woolley, S. C. and Howard, P. N. (2016). Automation, algorithms, and politics— political communication, computational propaganda, and autonomous agents—introduction. *International Journal of Communication*, 10:9.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion proceedings of the 2019 World Wide Web Conference*, pages 218–226.

Zarsky, T. Z. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.