UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Daniel Kansaon

The Dynamics of Coordinated and Malicious Behavior in Public Instant Messaging Groups

Belo Horizonte
2025

Daniel Kansaon

The Dynamics of Coordinated and Malicious Behavior in Public Instant Messaging Groups

Final Version

Dissertation proposal presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Fabrício Benevenuto de Souza

Belo Horizonte
2025

Resumo

A ascensão das mídias sociais transformou a comunicação e também criou novas formas de socialização. Nesse cenário digital em constante evolução, os aplicativos de mensagens representam uma parte importante, surgindo como poderosas ferramentas de comunicação pessoal e profissional. A imensa popularidade do WhatsApp, aliada à natureza da comunicação que ele facilita, criou um ambiente propício para diversos tipos de abusos online. Esta tese explora o ecossistema do WhatsApp, examinando as atividades e interações no submundo da plataforma. Ao revelar atividades ocultas e casos de uso indevido, o objetivo é obter uma compreensão mais profunda desse ambiente e identificar estratégias de mitigação. Com foco em grupos públicos, este estudo fornece uma análise detalhada do ecossistema da plataforma, destacando os abusos e seus impactos. Para isso, realizamos uma coleta de dados em larga escala, reunindo quase de 2 mil grupos políticos brasileiros e analisando cerca de 26 milhões de mensagens, incluindo textos, imagens, vídeos, áudios e figurinhas. Nosso estudo está estruturado em objetivos de pesquisa, cada um abordando diferentes perspectivas desse cenário complexo.

Primeiramente, direcionamos os esforços para observar a dinâmica dentro dos grupos, identificando abusos e comportamentos maliciosos. Nossas descobertas destacam dois principais tipos de ataques usados para interromper o funcionamento dos grupos no WhatsApp: o ataque de inundação (*flooding attack*), em que um invasor compartilha um grande número de mensagens, geralmente duplicadas, em um curto período de tempo, e o ataque de sequestro (*hijacking attack*), onde os invasores tentam obter controle total do grupo. Entre outras descobertas, identificamos que aproximadamente 7% dos grupos sofrem ataques de inundação, que geralmente são de curta duração (menos de quatro minutos), sendo que alguns grupos podem sofrer múltiplos ataques no mesmo dia. Além disso, verificamos que a maioria dos ataques de inundação (62%) são realizados usando figurinhas e que, na maioria dos casos, os invasores combinam ataques de inundação e sequestro para assumir o controle total dos grupos no WhatsApp. Nossos resultados enfatizam como os usuários em grupos públicos estão vulneráveis a diferentes tipos de ataques e conteúdos prejudiciais.

Em seguida, aprofundamos a observação de como formatos específicos de mídia podem ser explorados para fins maliciosos, com foco particular nos stickers. Embora os stickers sejam frequentemente percebidos como um meio inofensivo e lúdico de expressão, nossos resultados revelam que eles também podem ser estrategicamente utilizados para assediar, causar interrupções e disseminar conteúdo ofensivo em grupos públicos. Apre-

sentamos uma caracterização detalhada do uso de stickers em grupos políticos brasileiros, identificando padrões abusivos como ataques de flooding baseados em stickers, disseminação de imagens manipuladas ou enganosas e o direcionamento de indivíduos ou grupos específicos com conteúdo visual ofensivo. Essa análise evidencia o papel duplo dos stickers, que funcionam tanto como ferramenta de engajamento quanto como vetor de abuso, ressaltando a necessidade de estratégias de moderação mais sofisticadas, capazes de lidar com ataques multimodais e dependentes de contexto.

Por fim, focamos na observação da dinâmica das mensagens dentro dos grupos. Nessa direção, investigamos a existência de campanhas coordenadas no WhatsApp no Brasil. Utilizando análise de redes, nossos achados sugerem uma prevalência significativa de atividades orquestradas e coordenadas na propagação de notícias, sendo que 26% dessas mensagens têm origem em sites de desinformação. Além disso, descobrimos que as imagens desempenham um papel fundamental nas atividades coordenadas, representando 15% das mensagens, sendo frequentemente usadas para desinformação. Ademais, contas coordenadas também foram utilizadas para organizar ações coletivas, como os protestos contra os resultados das eleições.

Palavras-chave: Aplicativos de Mensagem, WhatsApp, Redes Sociais Online, Coordenação, Discurso de Ódio, Conteúdo Nocivo, Desinformação

Abstract

The rise of social media has transformed communication and also created new avenues for socialization. In this evolving digital landscape, messaging applications represent an important part, emerging as powerful personal and professional communication tools. The immense popularity of WhatsApp, coupled with the nature of the communication it facilitates, has created a fertile environment for diverse types of online abuses. This thesis explores the WhatsApp ecosystem, examining the underground activities and interactions within the platform. By uncovering hidden activities and instances of misuse, the goal is to gain a deeper understanding of this environment and identify strategies for mitigation. Focusing on the public groups, this study provides an in-depth analysis of its ecosystem, highlighting the abuses and their impact. To achieve this, we conducted a large-scale data collection, gathering almost 2k Brazilian political groups and analyzing over 26 million messages, including text, images, videos, audio, and stickers. Our study is structured into research objectives, each addressing different perspectives of this complex landscape.

First, we direct the efforts to observe the dynamics inside the groups, identifying abuses and malicious behaviors. Our findings highlight two primary types of attacks used to disrupt WhatsApp groups: *flooding attack*, where an attacker shares a large number of usually duplicate messages within a short period, and the *hijacking attack*, where attackers aim to obtain complete control of the group. Among other things, we find that approximately 7% of the groups receive flooding attacks, which are usually short-lived (usually less than four minutes), and groups can receive multiple flooding attacks. Also, we find that most flooding attacks are executed using stickers (62% of all flooding attacks) and that, in most cases, attackers use both flooding and hijacking attacks to obtain complete control of the WhatsApp groups. Our findings emphasize how users in public groups are vulnerable to different types of attacks and harmful content.

Second, we go deeper into observing how specific media formats can be exploited for malicious purposes, with a particular focus on stickers. While stickers are often perceived as harmless and playful means of expression, our findings reveal that they can also be strategically misused to harass, disrupt, and spread offensive content within public groups. We present a detailed characterization of sticker usage in Brazilian political groups, identifying abusive patterns such as sticker-based flooding, dissemination of manipulated or misleading imagery, and the targeting of specific individuals or groups with offensive visual content. This analysis highlights the dual role of stickers as both a tool for engagement and a vector for abuse, underscoring the need for more sophisticated

moderation strategies capable of handling multimodal and context-dependent attacks.

Finally, we focus on observing the message dynamics in the groups. In this direction, we investigate the existence of coordinated campaigns on WhatsApp in Brazil. Using network analysis, our findings suggest a significant prevalence of orchestrating and coordinated activity in message propagation of news, of which 26% comes from disinformation sites. Furthermore, we found that images are a key part of coordinated activity, comprising 15% of messages, which are also used to mislead. Moreover, coordinated accounts were also used to organize collective actions, such as the protests against the election results.

Keywords: Messaging Applications, WhatsApp, Online Social Networks, Coordination, Hate Speech, Harmful Content, Misinformation

List of Figures

List of Tables

Contents

Chapter 1

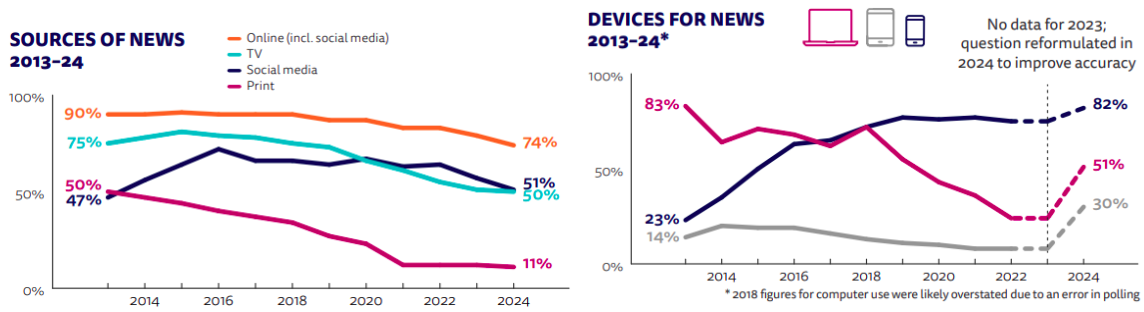
Introduction

Social networks have significantly transformed how people communicate and interact. Currently, more than a third of the world's population actively uses online social networks [36]. With the rapid growth of platforms such as Facebook, Twitter, and Instagram, individuals can stay connected with others across the globe, share their opinions, and access information instantaneously. These platforms have become an integral part of everyday life, shaping both personal relationships and professional networks. The influence of social media has redefined communication and also created new avenues for socialization.

In this evolving digital landscape, messaging applications represent an important part, emerging as powerful personal and professional communication tools. With billions of users worldwide, WhatsApp is the most popular messaging app, reaching around 2 billion monthly active users, surpassing other social networks such as Twitter and TikTok [140]. This is especially popular in Brazil, as everyone with a cell phone in Brazil uses WhatsApp on a daily basis [10], representing the second country with the most active users. As many mobile data plans do not charge for the data used through WhatsApp in Brazil, i.e., zero-rating policy, it became a very cheap means of communication and an alternative to SMS, voice, and video calls. This tool has seamlessly integrated into people's lives, becoming indispensable in various daily activities, including business proposes, event organization, entertainment, and news consumption. Moreover, a survey by Reuters observed an expressive growth in the proportion of people who used social networks for news consumption around the world. In Brazil, 51% of users said they consume news on social media, as shown in Figure 1.2(a).

The immense popularity of WhatsApp, coupled with the nature of the communication it facilitates, has created a highly convoluted and fertile environment for the propagation of misinformation campaigns. Since the presidential election in 2018, WhatsApp has evolved into a bustling media space for political militancy. This happens socially after the growing of use of mobile devices to consume news, as shown in Figure 1.2(b). The public space within the platform has emerged as a hub of communication and organization, enabling the seamless coordination of activists. These public groups make WhatsApp with the characteristics of social networks [126], intensifying the problems seen in other popular digital platforms due to anonymity and less transparency. A notable example of

Figure 1.1: Source people use to consume news.



(a) Sources of news consumption in Brazil. (b) Devices used to access news.

Source: Reuters Institute Digital News Report 2024 [106].

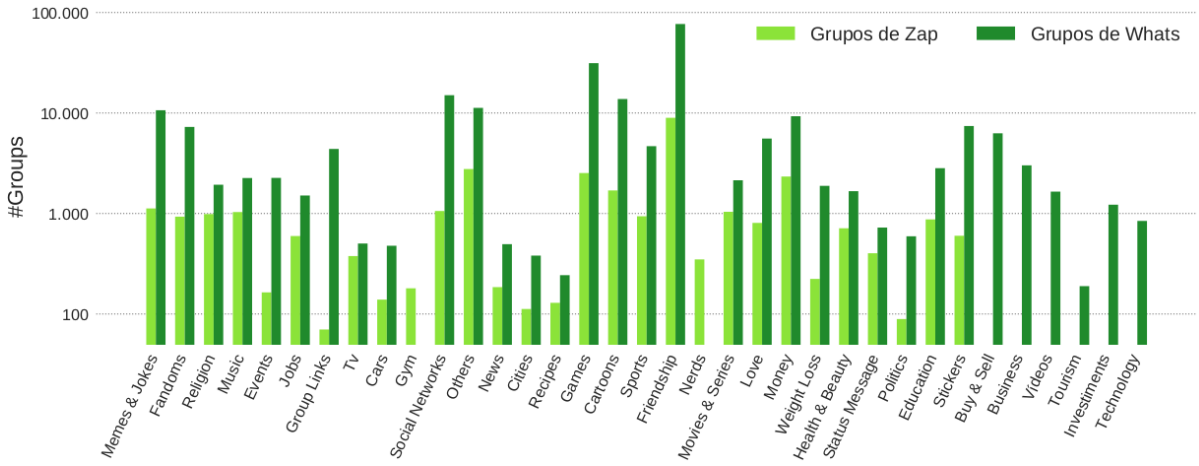
this phenomenon can be seen in India, where false reports of child abductions have led to numerous instances of mob violence and lynchings [135]. Also, in Brazil, a fact-checking effort conducted during the 2018 Brazilian presidential election process revealed an alarming statistic: 88% of the most popular messages circulating on the platform were found to be false or misleading [126]. More recently, in 2023, WhatsApp played a pivotal role in organizing and amplifying messages that called for protests, which led to riots and the invasion of the Congress and the Supreme Court in Brazil [42]. Moreover, prior research has shown that this lack of moderation facilitates various forms of harmful activity, including the spread of misinformation [126] coordinated political campaigns [149, 73], hate speech [133], and even criminal content such as scams and child abuse imagery [43]. In polarized political contexts, this unmoderated space has been increasingly exploited by political activism groups to manipulate narratives and attack opponents.

The environment created by messaging apps is quite complex. While these platforms offer a variety of tools to facilitate the rapid dissemination of messages, they also maintain a level of anonymity that conceals the authors of these messages. This is busted due to the end-to-end encryption structure that makes it hard to identify the origin of messages and track the content, and keep anonymity. Public groups on WhatsApp tend to form highly connected networks, reinforcing hierarchical information flows and echo chambers. These incidents have raised concerns about how such platforms are being used to influence public opinion, spread false information, and organize harmful activities.

Although public social networks have been extensively studied, WhatsApp presents unique challenges due to its private communication model, which differs from traditional social media platforms. This brings an important discussion about WhatsApp dynamics. The closed nature of WhatsApp makes it difficult to systematically monitor its internal dynamics, leading to interactions that remain hidden and fostering underground activities that are largely unknown. While prior research has exposed forms of abuse in other digital environments, there is still limited understanding of the structure of the WhatsApp ecosystem, how it is formed, how users interact within it, and what types of abusive or

manipulative practices unfold in these spaces. This thesis aims to explore the WhatsApp ecosystem, delving into the activities and interactions that occur within the platform. By investigating these underground activities and instances of abuse, providing a deeper understanding of the dynamics and offer a better comprehension of what exists in this environment.

Figure 1.2: Total number of groups by category in online repositories.



Source: The Author.

1.1 WhatsApp Groups Ecosystem

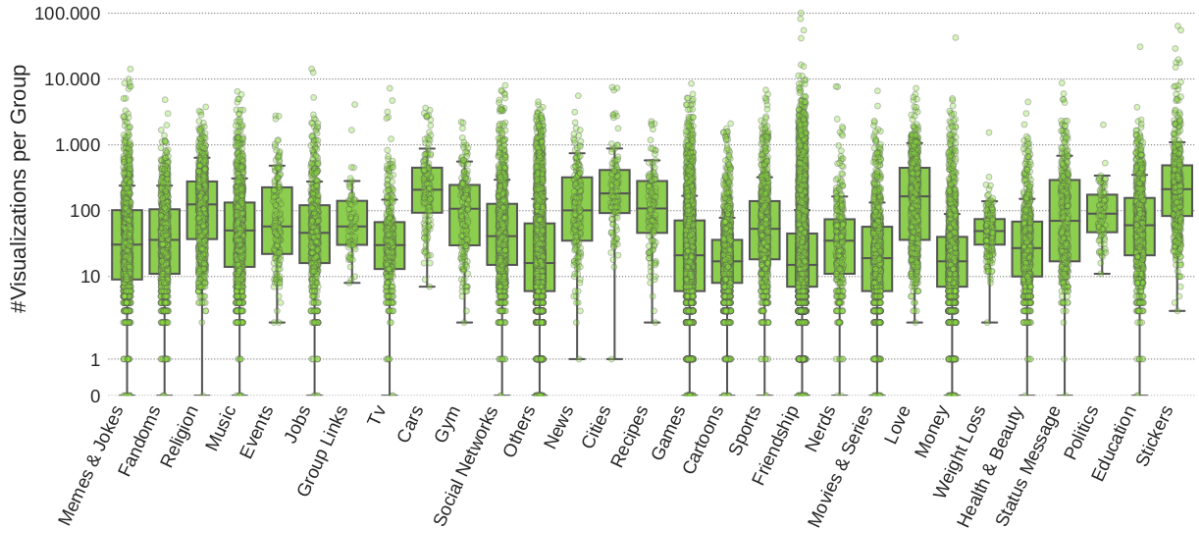
In order to understand the public ecosystem of WhatsApp groups, we analyzed the topic discussed in the existing public groups. Unlike other platforms, WhatsApp does not offer a native search interface for discovering public groups. As a result, users often rely on external repositories to find and join groups of interest. In this context, we collected over 240,000 group invite links from two major online repositories of groups: *Grupos de Zap* and *Grupos de Whats*. These platforms allow users to register public groups by submitting the group title, description, and selecting a predefined category.

Each repository includes category information provided by users during the group registration process. As shown in Figure 1.2, we identified 35 distinct topics of interest discussed in public groups on WhatsApp. Notably, "Friendship" was the most frequent category in both repositories. We also found groups include dating, meeting new people, and casual conversation. Although adult content is technically prohibited, many dating groups appear in this category and sometimes serve that purpose implicitly.

In addition, the *Grupos de Zap* repository provides the number of times each group page has been viewed. Since a single group can has many views, this metric helps

estimate the interest of each category. Figure 1.3 displays the distribution of group views by category. It is important to note that the most viewed categories do not necessarily align with the most frequent categories. For example, while the "Friendship" category has the highest number of registered groups, its median number of views is around 50. In contrast, other categories, such as "Romance", "Stickers", and "Cars", receive median views well above 200.

Figure 1.3: Distribution of group page views by category.



Source: The Author.

The "Stickers" group is particularly noteworthy. Despite comprising a relatively small number of groups, many sticker groups have accumulated over 10,000 views. These groups are often used to share personalized sticker packs, frequently based on popular memes or characters. Their popularity reflects a highly specific user behavior tied to WhatsApp's multimedia affordances.

Another notable finding involves political groups. Although politics represents a small fraction of the total number of groups (around 700), political groups receive a relatively high number of views, with a median close to 100. Given the centrality of political discourse in Brazilian society, this engagement highlights a disproportionate interest in political discussions within the WhatsApp ecosystem [102]. This observation motivates a deeper investigation into how political content circulates and how it is potentially weaponized through this platform.

In addition to online repositories, we also searched for groups on online platforms such as Twitter and Facebook, which are usually used to find and promote groups. In December 2021, we searched for public WhatsApp groups being advertised on social media. On Twitter, we used the official API[1] to collect all Portuguese-language posts with at least one WhatsApp group invitation link. For Facebook, we used the Crowdtangle

[1] https://developer.twitter.com/en/docs/twitter-api

API[2] to identify posts with public group links. After filtering, we retained approximately 20,000 groups from Twitter and 118,000 from Facebook.

To uncover the category of groups, we applied Latent Dirichlet Allocation (LDA)[3], a topic modeling technique widely used for uncovering latent semantic structures in text corpora [70]. We used both the text of the social media post and the WhatsApp group title as input for the model.

The results reveal a wide range of topics associated with WhatsApp groups shared on social media. On Facebook, we identified themes such as religion, self-help, commerce, and job opportunities, often involving product promotions and income. Twitter, on the other hand, featured a higher concentration of adult content, including pornography and the promotion of explicit material from platforms like Xvideos and OnlyFans. Additionally, we found more illicit content on Twitter, including groups advertising cloned credit cards, financial scams, and betting bots [75].

Despite WhatsApp's central role in Brazilian digital communication, little is known about what happens inside public groups. While external indicators suggest a wide variety of topics, ranging from entertainment and religion to commerce and even illicit activities, our analyses show that politics is one of the most prominent and engaging themes across different data sources. Furthermore, the platform's encrypted and closed nature makes it difficult to monitor the content that circulates, how it spreads, and who is behind its dissemination. Given its societal relevance and the growing concerns around misinformation, manipulation, and abuse in political discourse, this thesis focuses specifically on how politics is mobilized and weaponized within public WhatsApp groups, investigating its role in campaigns, content strategies, and coordinated behavior.

Although public WhatsApp groups cover a wide range of topics, including some with explicitly illicit content, political groups tend to attract the most attention. This is due to their high level of engagement, volume of messages, and their frequent involvement in controversial and impactful discussions. Despite the growing body of research on messaging application platforms, there remains a substantial gap in understanding the internal dynamics and abusive practices. Existing studies have predominantly focused on open social networks, where content is more accessible and subject to moderation mechanisms, leaving end-to-end encrypted environments underexplored. While prior work has mainly examined misinformation, few efforts have comprehensively analyzed the multimodal nature of abusive content within political contexts. Furthermore, there is limited understanding of how these different media types interact to amplify harmful narratives, and of the structural and temporal patterns of their dissemination. Given their prominence and societal impact, this thesis will focus primarily on political groups as a key lens through which to understand abusive dynamics in messaging platforms.

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

[3] We used the variational Bayesian implementation from the `scikit-learn` library.

1.2 Research Goals

The main objective of this thesis unfolds into the following specific research goals (RGs), each targeting a distinct aspect of political activity and abuse within public WhatsApp groups.

- **RG1 – Understanding attacks and hostile interactions in public groups.**

  To better understand the WhatsApp group ecosystem, it is essential to investigate the internal dynamics that occur within these groups. WhatsApp has become a significant platform for political activity, with public groups serving as hubs for communication, organization, and political coordination. However, while some research has explored the use of groups for misinformation and message propagation [102], the dynamics of these public groups, particularly how activist users join and act in pursuit of specific goals, remain largely unknown. In the polarized nature of political discussions, groups of activists organize themselves into digital militias to fight with each other and to promote hostile interactions toward opponents. This no man's land created within WhatsApp by digital militias is nearly unexplored by the research community. Therefore, this thesis aims to investigate how political activists organize within these groups and identify and characterize the hostile interactions by examining the dynamics and tactics used by these groups. In this direction, we focus on observing some important group features, such as the flow and frequency of messages, and key elements like group names, user messages, and posting frequency. In this context, we aim to address some key questions: *What types of attacks are common in political public groups?*, *How are these attacks executed, and what are their goals?* and *What kind of abusive content are users exposed to?*. Answering these questions will provide insights into how public spaces are abused.

- **RG2 – Understanding stickers as a new form of spreading harmful content.**

  While RG1 focuses on user interactions and attack strategies in political public groups, it is also crucial to investigate the media through which harmful content is disseminated. In this context, stickers have emerged as a powerful and understudied tool on WhatsApp. Originally designed to enhance expression in informal conversations, stickers are now widely used for political messaging, harassment, and even coordinated abuse. Their visual and multimodal nature allows users to encode offensive or ideological content in subtle ways. Despite their popularity, little is known about how stickers are used in political discourse, or how they differ from other media types in terms of impact and diffusion. Because WhatsApp allows users

to create and freely distribute custom stickers, this opens space for the proliferation of inappropriate, hateful, or misleading visual content. Therefore, this thesis aims to explore how stickers are used in political contexts, and how they can serve as vectors of abuse. In this direction, we seek to answer the following questions: *Do stickers share the same characteristics as other media formats on WhatsApp, and do users use them in the same way?*, *How is political content expressed through stickers in public WhatsApp groups in Brazil?*, *How do users abuse stickers to spread offensive content through public groups on WhatsApp?*.

- **RG3 – Identifying and characterizing coordinated strategies employed for message spreading.**

  Political activism groups often pursue diverse goals, with multiple groups collaborating toward a common goal. While individual groups may have limited members, users often participate in multiple groups simultaneously, creating connections similar to online social networks (OSNs) and potentially facilitating viral message dissemination [102]. While it is well-known that misinformation spreads virally within the WhatsApp ecosystem, the mechanisms behind this phenomenon remain unclear. Specifically, it is unclear how quickly some messages reach a large number of users and what the real impact is. Due to the difficulty of accessing and analyzing WhatsApp groups and their networks, understanding these dynamics is challenging. Therefore, this research goal aims to investigate message dynamics and uncover how messages spread widely within WhatsApp. Specifically, by investigating a large number of messages shared across multiple groups, this research goal aims to address the following research questions: *Is there evidence of coordinated accounts in the propagation of messages on WhatsApp?* and *What is the content and purpose of the coordinated messages? How are these messages related to recent political events in Brazil?*.

- **RG4 – Identifying subtle and harmful text-based discourse.**

  While the previous research goals focused on hostile user interactions (RG1), multimodal abuse (RG2), and coordinated message dissemination (RG3), another relevant dimension of harmful behavior in WhatsApp groups concerns the use of implicit discursive strategies to manipulate emotions and frame specific groups or individuals as threats. Building on our earlier findings regarding the use of stickers as a medium for harmful messaging, this research goal extends the investigation to text-based content, aiming to identify similarly subtle yet impactful discursive patterns. Despite increasing awareness of emotional manipulation in online environments, these indirect forms of harmful discourse remain understudied. Therefore, this goal is to propose a computational approach to characterize harmful textual strategies

in political discussions within WhatsApp groups. This involves addressing the following research questions: *How are politically harmful messages constructed?* and *What linguistic and narrative patterns are in such constructions?* By answering these questions, we aim to deepen the understanding of non-explicit mechanisms of manipulation and their role in shaping political conversations within WhatsApp groups.

Chapter 2

Background and Related Work

In this chapter, we provide a comprehensive summary of the background knowledge and related work essential for understanding this dissertation. We examine prior research efforts related to social networks and misinformation, particularly focusing on messaging platforms, with special emphasis on WhatsApp. Our discussion is structured according to the specific research goals outlined in Chapter 1, and we group relevant studies into thematic categories. Section 2.1, we review existing research on misinformation, exploring narrative analysis techniques to understand the creation and dissemination of false information. Section 2.2 examines coordinated and networked behaviors, highlighting strategies used in orchestrated online campaigns. Finally, Section 2.3 discusses previous studies on abuse and harmful content, emphasizing the detection and characterization of harmful and toxic behaviors within messaging platforms.

2.1 Misinformation and Narrative Analysis

We begin our discussion with a summary of previous studies of topics and narratives analysis, including misinformation spreading in the messaging apps ecosystem. In this context, WhatsApp has gained significant attention due to its widespread adoption and its prominent role in the dissemination of political narratives, often characterized by misinformation, conspiracy theories, and other forms of deceptive content. Public group chats, in particular, have been the focus of numerous investigations that examine how the platform is exploited for misinformation campaigns. Some studies have aimed to map and categorize the predominant themes discussed in WhatsApp groups, identifying categories such as religion, family, friendship, and news sharing [52, 75]. Within this context, politics consistently emerges as a dominant underlying theme across misinformation narratives [90]. Consequently, a growing body of research has examined these dynamics across a variety of national and sociopolitical contexts, highlighting both the global nature of the issue and the need for context analyses.

In Brazil, for example, studies have identified substantial evidence of misinformation campaigns, typically motivated by specific political or social objectives. [126] investigated the dissemination of information and misinformation in public WhatsApp groups during the 2018 presidential election. Their analysis revealed instances of false images being shared to foster narrative engagement. Similarly, [89] examined the circulation of misinformation during Brazil's 2018 presidential elections, raising concerns about political polarization and the potential incitement of violence.

Comparable phenomena have been observed in other national contexts, such as India. [76] evaluated the use of crowd-sourced tiplines during the 2019 Indian general election as a strategy to uncover misinformation. Their findings suggest that tiplines can offer valuable insights and serve as a promising tool for the early detection of misleading content. [145] conducted an interview-based study to explore how rural and urban communities in India interact with COVID-19-related misinformation on WhatsApp. Their findings demonstrate how factors such as class, urbanity, and social hierarchy shape the dynamics of misinformation, highlighting the need for context-sensitive to study this problem. In Indonesia, [109] examined who is exposed to political misinformation in the context of the 2019 election. The study finds that both traditional media (TV, newspapers) and social media platforms are positively associated with misinformation exposure, highlighting the misinformation on WhatsApp and Instagram.

Misinformation campaigns can have severe real-world consequences. India is an example that the spreading of rumors caused a series of violent lynchings around the country [6]. Also, in Brazil, there is evidence that fake news campaigns circulating on WhatsApp have interfered in the results of presidential elections [126, 17, 89]. During the COVID-19 pandemic, WhatsApp was also pointed out as an important vector for spreading fake news about health [69, 152]. Other issues regarding misinformation content disseminated through WhatsApp were reported in different locations, such as in India [76], in Indonesia [85], in Pakistan [69], U.K. [152], Ghana [105], Nigeria [23], Spain [41]. This shows us the misinformation is not a problem exclusive to one country, and the interconnection on WhatsApp allows the spread of messages and direct communications, which contributes to misinformation spreads.

During the COVID-19 pandemic, WhatsApp played a central role in the circulation of health-related misinformation, highlighting its influence not only within its own ecosystem but also across the broader social media landscape. The impact of misinformation dissemination varied across countries, contexts, and social groups. For instance, [152] investigated how age and message accuracy influence beliefs, credibility perceptions, and sharing intentions regarding COVID-19 misinformation on WhatsApp. The study finds that younger adults (18–54) were more likely to believe misinformation, while older adults (55+) showed signs of backfire effects after seeing corrections, highlighting the need for age-sensitive interventions. In the Brazilian context, [48] focused on the detection of

COVID-19 misinformation, proposing a method that achieved an F1-score of 0.778. Their results underscore the challenges of identifying misinformation in short-form messages like those in messaging applications. [69] analyzed user behavior around COVID-related content and quantified cross-platform information flow, revealing that WhatsApp often serves as a source for content that later appears on Twitter. [90] conduct a cross-cultural and multilingual analysis of COVID-19 misinformation, examining viral and debunked misinformation in English, Mandarin, and Farsi. The study revealed substantial variation influenced by culture, platform usage, freedom of expression, and government control. Similarly, [116] analyzed Twitter discourse in French, German, and Italian during the European vaccination campaigns, emphasizing the role of platforms such as Telegram and YouTube in amplifying low-credibility content.

Furthermore, studies show that misinformation on WhatsApp is not limited to textual messages [125, 19], multimedia content such as audio messages [94] and images [124, 53] have also been identified as carriers of misleading information. This concern is also perceived by users, who frequently cite WhatsApp as one of the platforms they are most worried about in terms of exposure to false or misleading content [107]. As a result, gaining a deeper understanding of how users engage with WhatsApp and what occurs within large public group chats operating under its end-to-end encrypted infrastructure has become an increasingly important area of research.

In parallel to misinformation studies, other research has examined the prominent role of messaging apps in the diffusion of this conspiracy theory. For instance, [65] conducted a large-scale, multilingual analysis of QAnon-related discourse on Telegram, investigating toxicity levels and thematic content across multiple languages. Their findings reveal that discussions revolve around far-right ideologies, global politics, COVID-19, and anti-vaccination narratives. Similarly, another study [5] investigated the dissemination of conspiratorial narratives by analyzing message forwarding dynamics, revealing that conspiracy discourse mixes a variety of narratives and alternative news sources, often entangled with legitimate media. [68] revealed financial incentives behind the spread of conspiratorial content on fringe platforms, in which channel operators monetize through e-commerce sales of dubious products, affiliate marketing, and crowdfunding. [141] presented a multimodal topic modeling study of conspiracy theories in German-language Telegram channels. Their findings show how textual and visual narratives intersect, indicating that conspiracy discourse is structured multimodally. In particular, memes, images, photos, and screenshots emerged as key components of conspiracy communication on Telegram.

Furthermore, other papers study narrative analysis of content that appears on messaging apps. [82] introduced a novel method for constructing topic networks that integrate both topic modeling and social media networks. They conducted a cross-platform analysis of the Nazi narrative pushed by Russian disinformation during the 2022 invasion

of Ukraine. They find that while the narrative was consistently present on Twitter, it only emerged on Telegram after the invasion. [56] proposed a real-time analysis with Hierarchical Agglomerative Clustering to examine the evolution of information narratives in pro-Russian and pro-Ukrainian Telegram channels during the initial months of the war in Ukraine. They find that narratives evolve distinctly and dynamically over time, reflecting substantial differences in perception and propagation mechanisms. [63] demonstrated how Russian media outlets utilized Telegram to maintain viewership and disseminate narratives. Their findings indicate that Telegram not only acts as a dissemination platform for state media content but also serves as a source of news itself-up to 26.7% of articles in some outlets referenced Telegram material.

Despite growing efforts to understand content dissemination in messaging applications, developing effective solutions to address disinformation remains a challenge. Several studies attempt to address these challenges through diverse approaches. Some propose classification models for identifying misinformation [48], while others focus on extracting features to better detect fake news [123]. Tools designed to support fact-checkers have also been introduced, such as platforms that monitor and display popular messages in real time [101, 72]. In terms of platform-level interventions, some research has proposed limiting message forwarding as a way to mitigate the viral spread of misinformation [102], a strategy later implemented by WhatsApp [103]. However, such measures have shown limited effectiveness, especially with the recent addition of communities. Complementary to these efforts, crowd-sourced tiplines have been explored as a method for early detection of misinformation [76]. Together, these studies underscore the multifaceted nature of the problem that is increasingly complex due to the private and encrypted nature of these platforms.

2.2 Coordination and Networked Behavior

With the increasing use of social media platforms, coordination has also become a fundamental component of online interactions. With the importance of speech and online interactions impacting the real world, many studies have shown interest in studying coordinated online behavior in recent years [93]. The use of bots is highly connected with misinformation campaigns, and they usually increase with political elections and economic discussions [30]. Coordination on social networks can influence how online interactions happen and even the user's perceptions, especially in misinformation campaigns, in which coordinated accounts often work together to boost false narratives [78]. Many studies highlight the prevalence of coordination efforts are an important part of misinformation

campaigns, which often lead to coordinated actions to maximize the reach of false narratives [78, 139, 146, 112] Furthermore, coordination is frequently used to amplify messages and manipulate trends [97, 168], contributing to echo chambers and online polarization [147].

The growing impact of online coordination on social media has motivated researchers to investigate its dynamics and to develop strategies for detecting and mitigating malicious activities [112, 119, 78]. Studies have shown that coordinated behaviors often play a central role in information manipulation campaigns. For example, a 2019 study reported that 71% of Twitter users mentioning trending U.S. stocks were likely bots, highlighting the prevalence of inauthentic activity in economic discussions [29]. Similar dynamics were observed during the COVID-19 pandemic, in which coordinated bots contributed to the spread of health-related misinformation [51]. In Brazil, this concern has led to discussions about legislation that would prohibit the use of bots on social networks [45]. Social bots have repeatedly played a strategic role in information operations ahead of major political events worldwide, reinforcing the importance of understanding and combating coordinated manipulation in online ecosystems [30].

The first studies that addressed the detection of automated accounts in online social networks focused on identifying spammers, uncovering structural differences between spam and legitimate users. For example, [164] revealed that spam accounts on Twitter exhibited distinct network characteristics, yet also highlighted the challenges of distinguishing spammers from legitimate users. Over time, new efforts emerged that leveraged supervised machine learning classifiers to identify bots based on behavioral and profile features [29]. While initially effective, these approaches quickly became insufficient as bots evolved. Newer bots are more similar to legitimate human-operated accounts than to other older bots, since they incorporate hybrid strategies [59]. As a result, individual-level classification proved increasingly ineffective. These traditional methods focused on identifying bots by analyzing individual user features [162, 163]. However, more recent studies have shown that modern bots often resemble human-operated accounts more than early bots, making their detection considerably more difficult [30].

Currently, the most promising advances in social bot detection are represented by group-based detectors, which focus on identifying suspiciously coordinated and synchronized behaviors rather than analyzing isolated accounts [30]. These approaches have gained attention, particularly after the introduction of the concept of coordinated inauthentic behavior (CIB) by Facebook [108], which emphasized the importance of detecting network group behavior rather than merely individual automation. Traditionally, the coordination studies focused on detecting automated accounts controlled by a script or program, designed to carry out specific tasks. The new definition focuses on coordinated actions at the group level that aim to achieve goals such as amplifying a message, manipulating public opinion, or spreading misinformation [93]. Unlike traditional methods that

estimate whether an account is a bot, this newer concept of identifying inauthentic coordinated accounts focuses on detecting coordinated actions at the group level by observing group users' patterns [112, 119]. Current approaches focus on identifying similarities in the action sequences of two or more accounts, modeling user activities to build user similarity networks, which makes it possible to identify coordinated user groups. This shift has led to the emergence of novel techniques capable of capturing synchronous activity patterns and behavioral similarities across networks of accounts.

Focusing on the role of coordinated accounts, a network-based framework has been developed by [112], to identify levels of similarity between coordinated accounts that connect users who post similar content. This user similarity network is built based on grouping coordinated actions. Additionally, several studies incorporate information about temporal user similarities to create the similarity network [118, 159, 119]. [119] focused on identifying coordinated accounts with synchronous activities, which means users who post messages simultaneously consider a window threshold limit that can be adjusted depending on the context.

Several studies across different social network platforms and political contexts have examined coordinated behaviors. [112] analyzed coordinated link-sharing behavior on Facebook during the 2018 Italian general election, uncovering networks of pages, groups, and public profiles that shared the same political news links within narrow time frames, showing evidence of coordinated inauthentic behavior. [153] studied Twitter activity leading up to the January 6th U.S. Capitol attack and identified coordinated strategies to propagate misleading claims about the 2020 U.S. election. Similarly, [14] analyzed tweets related to the 2017 French presidential election, revealing that some accounts heavily retweet one another, becoming especially active around critical political events and promoting their preferred candidates at significantly higher rates than organic users. [79] extended this line of inquiry to YouTube, conducting an exploratory study using rule-based classification and network feature engineering. Their findings indicate that suspicious channels exhibit more intense anomalous activity peaks and can be clustered based on coordination patterns.

Some studies have also focused specifically on detecting coordinated behavior in messaging platforms. These studies typically construct similarity networks, where user accounts are connected by an edge if they post identical or highly similar content. The weight of each edge often reflects the frequency or intensity of this similarity. Based on definitions of group-level coordination, a set of users is considered coordinated if they exceed a minimum threshold of co-occurrence. This threshold can be determined using various techniques, such as the Disparity Filter algorithm [149], grid search approaches that explore different combinations of edge weight and centrality metrics [25], or predefined fixed thresholds [110].

In this context, [110] investigated the spread of disinformation on Telegram during

the COVID-19 pandemic, revealing that bots dominated conversational activity, while human users played a critical role in amplifying and forwarding the content. [3] found that bots were strategically deployed within the Islamic State's online ecosystem on Telegram to amplify propaganda, sustain user engagement, and minimize operational risks. Similarly, another study [25] examined coordinated inauthentic behavior related to the 2024 U.S. election across Telegram, Facebook, and Twitter, finding that highly partisan and conspiratorial content spreads through tightly coordinated communities. In a related context, [149] analyzed political mobilization efforts on Telegram during Brazil's 2022 presidential election, demonstrating the platform's strategic use in orchestrating influence operations during key political moments.

On WhatsApp, [115] presented a hierarchical network-based analysis of user participation, focusing on content and the network structure formed by users sharing identical content. Their findings show that even in a platform limited to small-group communication, users can build broad dissemination networks. Similarly, [114] analyzed political information dissemination on WhatsApp by constructing a media-centric network connecting users who shared the same content across different public groups. Using community detection and temporal analysis, the study identifies dynamic user communities that transcend group boundaries and play a central role in spreading information. [73] investigated political WhatsApp groups during Brazil's 2022 election and found that coordinated dissemination efforts focused on promoting news content linked to misinformation sources. The study also revealed that coordinated accounts played a key role in organizing protests and attacks aimed at contesting the election results.

2.3 Abuse and Harmful Content

The study of online speech, particularly on online social networks, has become a significant area of research in recent years. The anonymity provided by these platforms increases the likelihood of aggressive behavior and encourages online expression [15], which contributes to the prevalence of toxic [71], abusive [113], discriminatory [142], extremist [113], and hateful [49] comments. The growing research interest is driven by the significant impact of such online speech, particularly its role in attacks against minority groups, raising serious social and ethical concerns. Detecting and removing such speech is also a concern for platforms that aim to have a less toxic environment. A toxic environment negatively impacts user engagement, and since these platforms rely on advertisers, this issue is also relevant to them. To address this, platforms have policies to define and regulate what is allowed, deleting or suspending hateful posts and users. For example, in Facebook and

Instagram terms, hate speech is defined as *"violent or dehumanizing speech, statements of inferiority, calls for exclusion or segregation based on protected characteristics, or slurs* [104]". Twitter prohibits targeting individuals or groups with content referencing forms of violence [47]. Similarly, YouTube does not allow any content that promotes violence or hatred against individuals or groups [166].

Abusive language refers to aggressive, harmful, or threatening language, and it includes hate speech, derogatory language, and profanity [113, 49]. Within this context, various definitions categorize harmful language into broader categories, such as hate speech [49], aggressive speech [22, 50], and offensive language [31]. In this section, we detail two critical forms of harmful speech: hate speech and fear speech.

2.3.1 Hate Speech

Social media platforms allow the spread of hate speech, mainly due to anonymity, which encourages more aggressive behavior among users [15]. In this sense, hate speech is an online manifestation of societal conflicts, and it is widely acknowledged as a significant issue by societal groups and governments [144]. However, a key challenge remains the lack of a clear definition of hate speech [49]. Even in our society, there is no uniquely accepted definition, and many works use overlapping terms such as abusive, toxic, dangerous, offensive, or aggressive language [121]. Furthermore, hate speech moderation raises legal concerns, particularly regarding freedom of opinion and expression [144].

One definition of hate speech is described by [49] as *"language that attacks or diminishes, incites violence or hate against groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or other traits"*. The spread of hate online has a significant societal impact, resulting in harm to individuals and communities, which is boosted by social networks due to their viral nature and user anonymity. This raises questions about moderating online spaces. As a result, many studies have emerged proposing solutions to moderate this phenomenon online, focusing on detecting and characterizing hate speech [169, 96].

To tackle the challenge of identifying and mitigating hate speech, the research community has made significant progress through the development of curated datasets and the application of machine learning models. One of the most widely adopted strategies involves constructing lexicons with hate terms, which serve as foundational resources for both detection and analysis tasks. Notable among these is the HateBase lexicon [62, 64], a multilingual repository designed to catalog hate speech across various cultural and linguistic contexts. Other efforts include the development of domain-specific lexicons

targeting particular communities or forms of abuse. These lexicons are useful not only for detection systems but are also frequently used to enhance the feature representations in supervised learning approaches [134]. In this direction, various lexicons have been developed for specific contexts, such as anti-Jewish [156] and anti-Muslim hate [151], as well as for tasks like hate speech and offensive content classification in social media, as addressed by the HASOC shared task [31, 92].

Another research direction has explored machine learning models, such as Logistic Regression and SVM [88], leveraging traditional text features, including TF-IDF, word embeddings, n-grams, and POS tags to identify hate speech [31, 83, 157]. Despite their relative simplicity, such approaches have shown competitive performance on benchmark datasets and played a foundational role in early automated hate speech detection systems. More recent studies have explored deep learning approaches. For example, [127] proposed Convolutional Neural Networks (CNN) to predict whether a tweet with a given target is hateful or not hateful, and whether the hate speech is directed at a specific person or a group of individuals. [148] proposed Long Short-Term Memory (LSTM) models to classify messages from six publicly available datasets into three classes, abusive, hateful, or neither. Furthermore, with the rise of transformer-based language models, pre-trained generative transformer models have gained attention for their effectiveness in detecting hate speech [8, 4, 161]. [4] performed the first extensive evaluation of multilingual hate speech detection, analyzing the performance of different deep learning models in various scenarios. They observed that BERT-based models perform much better in high-resource scenarios. Furthermore, they also observe that simple techniques such as translating to English and using BERT achieve competitive results in several languages.

In the other direction, a more recently adopted approach focuses on toxicity detection. The Perspective API[1] is a machine learning model that evaluates text based on a toxicity index, identifying comments that are rude, disrespectful, or likely to make participants leave a discussion. This machine learning model was developed through supervised learning, using a vast dataset with comments gathered from diverse online platforms, encompassing more than 20 languages [117]. Since toxicity encompasses offensive, abusive, and aggressive content, this approach simplifies detection. As a result, many studies have leveraged the Perspective model for hate speech detection [65, 138].

Most studies in this field focus on binary classification, distinguishing between hate speech and non-hate speech messages [121]. Others explore explicit and implicit hate speech to identify more subtle forms of hateful content [158, 40]. In addition, some studies focus on the detection of accounts that consistently produce harmful content [128, 95]. Overall, research in the area of hate speech has made significant advances, mainly in hate detection techniques, especially driven by deep learning techniques and transformer-based models [1], which allow for context-aware language understanding and transfer learning

[1] https://perspectiveapi.com/

across domains. However, the architecture of online platforms still presents significant challenges, and the differences in platform structures and the complexities of regulations make it difficult. A growing body of research has raised concerns regarding biases in hate speech and toxicity detection models, particularly in multilingual settings. Despite the popularity of Perspective API, recent findings reveal that it systematically assigns higher toxicity scores for some specific language content compared to equivalent content in other languages, such as English [117]. These challenges are further amplified in the context of messaging applications, where limited context, user privacy, encryption, and multilingual communication environments make bias detection and mitigation even more complex.

2.3.2 Fear Speech

While substantial progress has been made in the detection and classification of hate and abusive speech, more subtle and insidious forms of harmful content, such as fear speech, remain relatively underexplored in the literature [132]. Although hate speech is often the focus of content moderation systems due to its overtly aggressive or dehumanizing nature, recent research suggests that a significant fraction of harmful discourse in online platforms is driven not by hate, but by fear speech [80].

Fear speech is characterized not by explicit hostility or insults, but by the strategic manipulation of emotions to instill (existential) a sense of fear in the mind of readers of another group [18]. This type of speech creates narrative, often relying on misinformation and conspiracy theories about one or more online communities [18, 133]. Unlike hate speech, which tends to be more visible due to the use of violent language, fear speech typically adopts a more persuasive, and even fact-like structure, making it more difficult to detect using conventional toxicity or hate classifiers. This subtlety presents major challenges for automatic detection. Empirical studies have shown that traditional classifiers trained on hate speech often fail to distinguish fear messages from political discourse, particularly because fear-inducing messages may not contain overtly toxic terms. For instance, [132] presents the following example of fear speech, which does not rely on profanity or slurs.

"Hundreds of South Americans are marching through Mexico, aiming to cross the US Border and demand asylum in the US. No one in Mexico is stopping them. This is a national security threat and should be dealt with by force if necessary. What else is our military good for if they can't stop an invading force?".

Although this message does not contain toxic language, it presents arguments

supported by evidence and spreads fear, urging users to take action, and threat to national security, all while maintaining a tone that appears rational and fact-based. As a result, fear speech is a powerful mechanism for reinforcing polarization, promoting discrimination, and contributing to radicalization within digital communities [133].

To address this challenge, some works emerged focusing on understanding and quantifying this type of speech. [18] examined the boundaries of freedom of expression in contexts where speech may contribute to violent conflict, proposing the concept of fear speech. With this definition, later, efforts have been directed to fear speech characterization and automatic detection. In this context, [133] conducted the first large-scale study of fear speech on WhatsApp groups. This study revealed that fear speech messages spread faster than traditional toxic content and are harder to detect with existing NLP models due to their subtle and low-toxicity nature. Additionally, users who post fear tend to be influential and occupy central network positions, contributing to the popularity of such messages. The study also observed that many fear messages contain misleading information. While the authors developed a model (XLM-Roberta with Logistic Regression) to classify and identify fear speech, it achieved only 0.51 precision in the fear class, highlighting the challenges of classifying these messages. A more recent study analyzed fear speech on Gab.com, comparing the influence and reach of users posting fear speech [132]. This research confirmed that users who frequently post fear speech tend to accumulate more followers and hold more central network positions than those who post hate speech. Furthermore, with a BERT model, they achieved a macro F1-score of 0.62 for detecting fear speech.

Recent efforts have extended the study of abusive and fear speech to messaging platforms, which present distinct technical and sociological challenges due to their ephemeral, private, and encrypted nature. Among these, Telegram has emerged as a particularly relevant case, given its widespread use by fringe communities and political movements. In this context, [160] proposed a comprehensive framework for detecting abusive language and identifying hateful communities on Telegram. Their architecture includes a message-level classifier that detects abusive language using supervised learning techniques and a channel-level hatefulness classifier that integrates topic modeling to characterize the toxicity of entire communities. This multi-level approach acknowledges the collective and contextual nature of abuse, where harmful narratives are not merely the sum of individual messages but are constructed and amplified within networked group dynamics.

In a complementary line of investigation, [150] conducted a longitudinal study on the temporal evolution of hate speech in an Italian conspiracy-themed Telegram channel during the first year of the COVID-19 pandemic. The study revealed that the targets of hate speech were not static and shifted over time in response to broader sociopolitical changes. These findings highlight that hate speech on messaging platforms is highly

reactive to external events. In parallel, a new line of research has also focused on detecting "othering" language, which is defined as the process of portraying an outgroup as fundamentally different and inferior [2]. This approach provides a theoretical grounding that bridges sociolinguistics and computational modeling. [2] proposed a novel approach to cyber hate speech detection by identifying linguistic patterns associated with othering messages. They developed a specialized feature set and employed embedding-based models to learn semantic relationships underlying hateful narratives.

Building on these insights, [57] proposed an advanced computational framework that combines sociological theory with large language models (LLMs) to identify othering language in content from fringe platforms such as Gab and Telegram. Their model integrates contextual embeddings derived from LLMs with socio-discursive features inspired by theories of intergroup conflict. Focusing on the Russia–Ukraine war, they demonstrated that othering intensifies in times of geopolitical crisis and often coincides with moralized, emotionally charged narratives that frame the outgroup as an existential threat. The proposed model not only captured these complex rhetorical dynamics but also outperformed traditional hate and fear speech detectors, achieving an F1-score of 0.77. This result suggests that incorporating sociological insights into computational pipelines can significantly enhance the detection of nuanced and emergent forms of harmful speech.

Despite recent advances, detecting fear speech remains a significant and unresolved challenge. Its subtle rhetorical structure and reliance on emotional manipulation rather than hateful language make it difficult to identify using conventional moderation tools or existing toxicity classifiers. Unlike hate speech, which often includes terms that violate platform policies, fear speech frequently contains persuasive discourse, allowing it to escape automated detection and human moderation. Moreover, most existing research has concentrated on textual manifestations of fear speech, while ignoring the presence of multimodal content, such as images, videos, audio, and especially stickers, forms of media that are prevalent in private messaging environments. Additionally, fear speech is highly context-dependent, varying in narrative structure, targets, and cultural cues across regions and crises. For instance, [133] documented fear-based narratives centered around Muslim communities in the Indian electoral context, yet similar rhetorical strategies may take on distinct forms in other sociopolitical settings. This variation underscores the need for cross-contextual and culturally sensitive approaches, as well as the development of generalizable models for capturing fear messages across different sociocultural landscapes.

Chapter 3

Data Collection

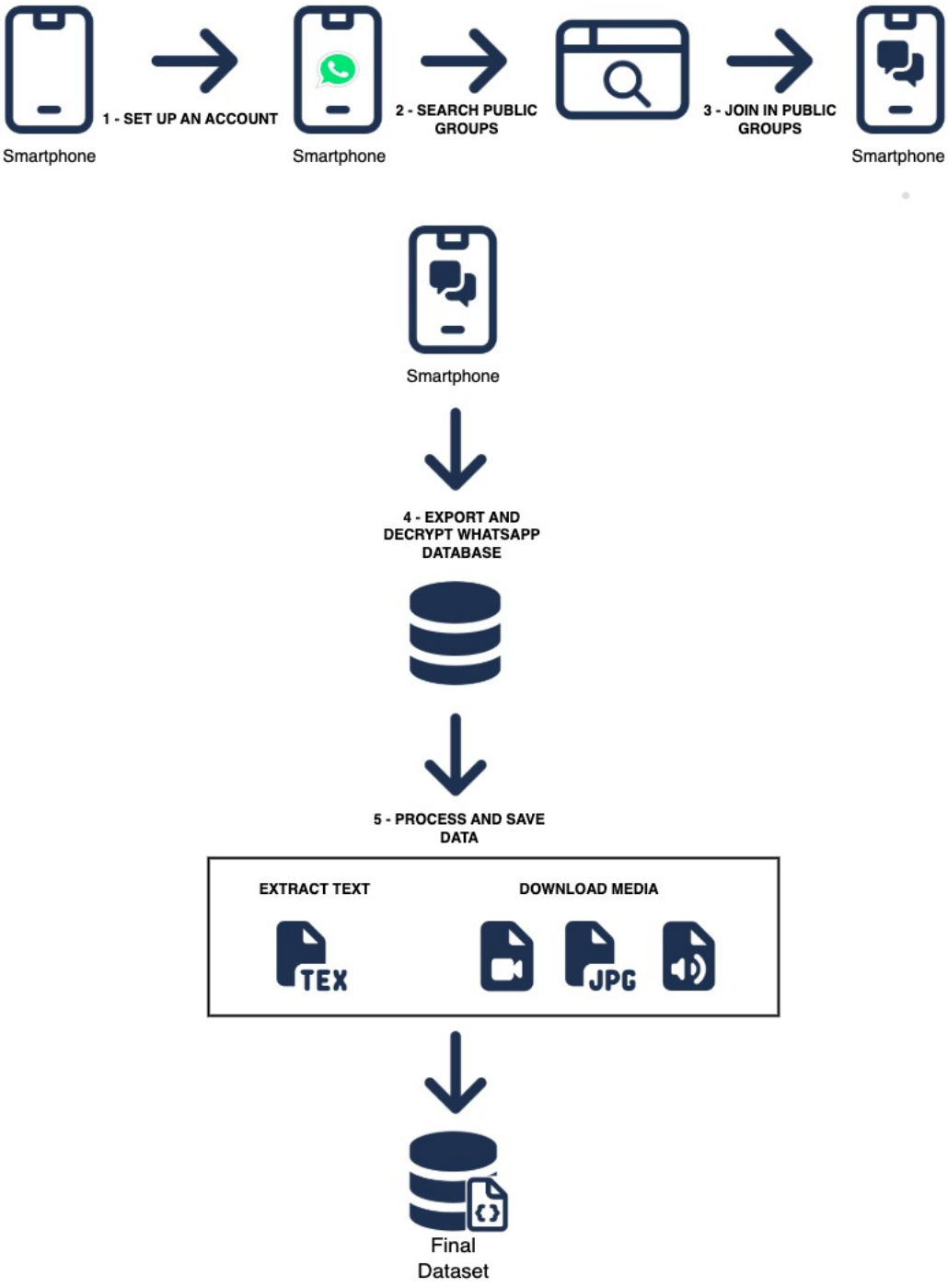
Although WhatsApp's public ecosystem shares some similarities with other online social networks (OSNs), its data accessibility is very different. To participate and interact in groups on WhatsApp, users must have a smartphone with a valid account to participate and interact in groups. This adds an extra layer of complexity, necessitating managing physical devices. The differences between WhatsApp and other apps extend to architecture and data access. Unlike platforms like Twitter, where most interactions are public, WhatsApp operates primarily through private groups and encrypted one-on-one chats. Additionally, there is no structure to search for specific interest groups, making it challenging to discover groups and interests to collect.

The common approaches used to explore and study WhatsApp typically focus on public space and public groups, where access is possible via an invite link. Currently, studies that explore WhatsApp data use two main strategies. The first involves scraping data through the web interface with WhatsApp Web, while the second approach, which is used in this work, consists of directly decrypting the WhatsApp database from the smartphone.

Although the web-based method appears simpler and commonly used in recent studies [33], it contains some limitations. Access to data is often restricted to what is visually displayed on the interface. Moreover, WhatsApp's frequent updates to its web interface can disrupt the data collection process, requiring continuous adjustments to the scraping tools. Additionally, large-scale data collection through the web can lead to slow connections to the interface, further limiting the data collection process. In another perspective, the decryption method used in this study was first proposed by [54]. It involves rooting the phones to have access to the WhatsApp dataset and retrieving the encryption key, which is then used to decrypt the messages directly. This approach enables the complete access to the entire dataset.

In this chapter, we will outline the entire process used to collect WhatsApp data, starting from setting up an account, searching and joining public groups, decrypting the WhatsApp dataset, and processing the data, as shown in Figure 3.1.

Figure 3.1: Methodology Steps for Collecting WhatsApp Data.



Source: The Author.

3.1 Set up of an Account

To collect data from WhatsApp, it is necessary to join the groups and participate on the same level as other users. Unlike other social media platforms such as Telegram, WhatsApp does not provide an API or direct access. For this, a physical phone with

the WhatsApp app properly installed and configured is essential. As a result, the data collection process involves creating accounts and setting up WhatsApp on the phone. In this context, it is important to ensure that the account configuration complies with WhatsApp's Terms of Service and that group monitoring is conducted ethically. It is important to mention that, although we are on the same level as other users, we do not interact in the groups or with any participants. Our purpose here is only to create accounts for monitoring and collecting messages from WhatsApp groups.

In this direction, the first step in the setup process involves downloading and installing WhatsApp on the phone. To register an account, users are required to accept the Privacy Policy and Terms of Service. Subsequently, a valid mobile number is needed, as the WhatsApp app requires it for verification by sending an SMS with a code for authentication. This highlights an important difference, as unlike web-based platforms, WhatsApp accounts are associated with a single mobile number. Finally, the app is authorized for use after receiving and validating the authentication code.

After accessing the app, the next step is adding personal information to complete the account setup. WhatsApp restricts users to a single account per phone number, so to manage multiple accounts, we created three distinct profiles. To simulate realistic user profiles, we used the persona concept, based on user surveys. In our study, the persona is used to create synthetic user profiles, as it effectively captures the characteristics and behaviors of typical users of the application.

To build a persona, we first define personality traits that align with the intended use of the app and the groups we aim to monitor. For this purpose, we adopted the persona definition framework used in previous research exploring the same political context in Brazil on WhatsApp [33]. The defined persona is as follows: *"A woman, approximately 40 years old. She uses WhatsApp to talk to her family, friends, and she gets information by following groups of political content, although she does not send messages to these groups. She uses WhatsApp exclusively from her cell phone to chat with others. She shares the news and reports she finds relevant, and her group partners often include her in new groups."* [33]. In addition to creating the persona, we specify in the WhatsApp profile status that the account is intended for group research purposes.

3.2 Search and Join in Public Groups

After configuring the phone and the account, the next crucial step is to find and define which public groups we will monitor. Since WhatsApp does not offer an official mechanism for searching groups by subject or name, the most appropriate approach is to repli-

Figure 3.2: Example of user Sharing Group Invitation.

Source: The Author.

cate the steps taken by users to discover groups. Public group creators typically aim to attract as many participants as possible and to achieve this, they often share invitation links on social media platforms. Thus, users can join these groups by clicking on the shared links and becoming part of the community. This invitation URL for a WhatsApp public group follows a standard template, such as "*https://chat.whatsapp.com/<identifier>*". By clicking on this link, the user is redirected to the group to confirm whether they wish to join. Figure 3.2 illustrates an example where a Twitter user shares an invitation link for a group. Typically, the invitation contains a call to action encouraging others to join, followed by the invitation link.

By searching for invitation URLs on social media platforms and search engines, it is possible to find groups on a variety of topics. However, in our case, we focus on identifying political groups that discuss topics relevant to the Brazilian political scenario. To achieve this, a commonly employed strategy, as demonstrated in prior works [54, 126], involves using keywords to search for invitations associated with the selected topic.

For group discovery, we used a set of keywords related to the Brazilian political scenario, initially proposed by [126]. Additionally, we expanded this list with updated terms reflecting relevant figures and topics that gained relevance after 2018 (see Appendix A), when the initial list was created. Using this comprehensive keyword list, we conducted searches across social networks such as Twitter and Facebook, search engines like Google, and online repositories indexed by Google. These repositories included platforms like *https://gruposwhats.app*, *https://www.gruposdewhatss.com.br*, and *https://gruposdezap.com*."

By searching with these keywords, we collected a set of URLs for public WhatsApp groups. Using the links, we can gradually join the groups. Sometimes, these invitations can be revoked by the group admin, and some of the collected invitations can be not available to join, as we did not join the links in real-time when they were posted. Despite this, most of the links remained available for an extended period, allowing us to build a solid dataset of groups. Besides the fact that the process of joining groups can be auto-

mated, automatically joining them could cause WhatsApp to detect suspicious behavior, potentially leading to a ban from the platform. For this reason, we decided to join groups manually to mitigate this risk.

3.3 Export and Decrypt the WhatsApp Dataset

Once we join all the selected groups, the phone will receive all messages and notifications from those groups. However, the data remains only on the local device, and we need to access this local dataset to retrieve, process, and analyze the messages. There are two common approaches to this. The first is the web-based method, which is limited by the data accessible through the interface, typically restricting it to only what is visually displayed. Furthermore, frequent updates to WhatsApp's web interface can disrupt data collection, and large-scale data retrieval can result in slow connections. The second approach, proposed by [54], involves rooting the phone to access the WhatsApp dataset and retrieve the encryption key, which is then used to decrypt the dataset. Although this method provides complete access to the entire dataset, it requires more configuration steps to access the key to decrypt the messages.

Due to the limitations of the web-based method, our study used the rooting approach proposed by [54] to collect messages. The first step is to export the WhatsApp dataset with the messages received in groups. However, WhatsApp data is stored on the device in an encrypted database, and the challenge lies in decrypting and extracting the messages from the dataset. WhatsApp uses end-to-end encryption, ensuring only communicating persons can read or listen to what is sent. Therefore, accessing these messages requires decrypting the dataset, which is only possible with the private key. The decryption process is easier when the phone is rooted, as in our case, where we used three Samsung Galaxy A30S devices running the Android operating system. WhatsApp stores the private key in the system storage, accessible by navigating to the location `/data/data/com.whatsapp/files/key/` on the phone. Accessing this path, we then extract and save the key to decrypt the dataset. To achieve this, we used the WhatsApp-Crypt14-Decrypter[1], which takes both the key and the encrypted database as input, processes them, and saves the decrypted database as a file named `msgstore.db`. This file is a SQLite3[2] database, a C-language library that implements an SQL database engine and can be manipulated programmatically.

With the decrypted dataset, we can easily manipulate the received messages from

[1] https://github.com/pharaoh1/WhatsApp-Crypt14-Decrypter
[2] https://www.sqlite.org/

groups by executing `SQL SELECT` commands to extract information.

3.3.1 Dataset Structure

After decrypting the dataset, we can access to the `msgstore.db`, an SQLite3 database. This dataset contains all crucial information related to the WhatsApp account of the extracted phone. The database is structured into several tables, each storing different data types essential for the WhatsApp app. However, not all tables are relevant to our analysis. The following are the most important tables:

- **Message**: This table contains the messages sent on groups and their metadata. It includes `chat_row_id`, which links each message to a specific chat by referencing the `Chat Table`; `_id`, which serves as a unique identifier for each message; `sender_jid_row_id`, which references the sender's identifier in the `Jid Table`; `timestamp`, a timestamp indicating when the message was sent; `text_data`, which stores the content of the text message (empty for media messages); and `message_type`, which specifies the type of message (e.g., text, image, video).
- **Chat**: The `chat Table` contains metadata about all chats. This table stores fields such as `_id`, which is the primary key identifying each chat; `subject`, which represents the title or name of the chat (e.g., a group name).
- **Jid**: This table contains information about all users who are part of groups.
- **Message_Quoted**: This table has relationships between messages that quote or reply to other messages. It contains `message_row_id`, which links the quoting or replying message to the corresponding record in the `Message Table`, and `parent_message_row_id`, which identifies the original message being quoted or replied to, and `parent_message_chat_row_id`, which identifies the chat in which the original (parent) message exists.
- **Message_Forwarded**: This table stores information about messages that have been forwarded. It contains `message_row_id`, which links the forwarded message to the corresponding record in the `Message Table`, and `forward_score`, which indicates how often the message has been forwarded.
- **Message_Ephemeral**: This table stores information about ephemeral messages, including their duration and expiration timestamp. It includes fields such as `message_row_id`, which links the ephemeral message to the corresponding record in the

`Message Table`; and `expire_timestamp`, which records the exact time when the message will expire.

- **Message_Media**: Contains metadata about message media, such as images, videos, and documents. It includes `message_row_id`, which links the media file to the corresponding record in the `message` table; `media_key`, which stores the encryption key for the media file; `media_key_timestamp`, which records when the media key was generated; `direct_path`, which indicates the server path to the media file; `message_url`, which stores the URL for downloading the media; `mime_type`, which identifies the type of media (e.g., image/jpeg); `file_size`, which records the size of the media file in bytes; and `file_hash`, which ensures file integrity through a hash value.

3.4 Process and Save Data

Once the WhatsApp database is decrypted, we gain full access to its raw data, allowing us to manipulate and analyze it. By joining the data from various tables within the database (See Section 3.3.1), we can extract and correlate crucial information, such as message content, timestamps, sender, and group interactions. To ensure privacy, we anonymize all personally information, including phone numbers and any details that could reveal the identity of group members.

Since all messages are stored locally on the smartphone, we periodically run processes to extract this data and download the media, typically every 7 days. Media is typically available for up to 15 days on WhatsApp, which is sufficient for our purposes. During this process, we calculate metrics such as the total number of media appearances, repeated content, and other relevant details.

By examining the Message table, we can access all WhatsApp messages, each identified by a unique identifier generated by WhatsApp and stored in the local dataset. Additionally, we can retrieve information about the sender, the timestamp of when the message was sent, and the message type (e.g., text, video, audio, image, document, or sticker). By joining the Message table with the Chat table using the `chat_row_id`, we can access details about the group where the message was sent. This includes the group's unique ID, title, and creation date. Moreover, we also get if the message was forwarded or not, which helps us understand how the message was shared in the group, which can be useful for analyzing the message's dissemination and potential virality. Finally, in the Message Quoted Table, we can determine if the message was quoted or replied to.

Figure 3.3: JSON Structure of WhatsApp Message After Processing


```
{
  "media": {
    "hashes": {
      "checksum": "1ab36247943c743d78f21a8c6fea04d7",
      "phash": "bf17437060434b3f"
    },
    "width": 1035,
    "height": 1600,
    "stored_filename": "Av5qbQPb24.jpeg",
    "media_key_timestamp": 1672182612000,
    "media_type": "image",
    "file_length": 93288,
    "partial_media_hash": "ZWYbVfikNDiJr8XwTKuH0H9Spx=",
    "message_url": "https://mmg.whatsapp.net/d/f/Av5qbQPb24.enc",
    "media_key_date": "2022-12-27 23:10:12",
    "mime_type": "image/jpeg"
  },
  "group": {
    "group_ddd": 61,
    "group_creator": "57c2fac85bdba0c4433983bd0a221d47",
    "group_country": "BR",
    "group_date": "2018-10-04 01:30:43",
    "group_id": "60300ef08ece36f87964aa4b1545f1ea",
    "group_name": "PATRIA AMADA BRASIL"
  },
  "sender": "e2c32e15b0775768291d95fa5851283c",
  "is_quote": false,
  "timestamp": 1672185714000,
  "sender_id": 432840,
  "ddd_code": 62,
  "country_code": "BR",
  "text": "Olha essa bandeira.",
  "date": "2022-12-28 00:01:54",
  "row_id": 88,
  "message_type": 1,
  "message_id": "3A1850A332B214ADFFC7",
  "forwarded": 1
}
```


Source: The Author.

3.4.1 Extract Text and Download Media

With this initial data processing, we gain an overview of all messages. However, we must further investigate the content of each message. This process is divided into some steps. First, we collect textual messages, as they are readily accessible in their raw

form in the database. Next, we download and process media messages, including images, videos, documents, and audio files. This step is crucial because physical media files are not directly stored in the database. Instead, we can access the media from WhatsApp's servers using the message URL. To retrieve the content, we use the media URL along with the corresponding media key to download the media. The media file is stored in an encrypted format on WhatsApp's server. To access it, we send a request to the server, which allows us to download the file. Then, we decrypt the file using the corresponding media key, enabling us to save the media locally for further processing.

After downloading the media and gaining access to all messages and their associated metadata, it is essential to collect the message content. Although the WhatsApp database contains message metadata, each message is stored as a distinct object. This means that when the same message is shared across multiple groups, it is saved as separate instances in the dataset, making it difficult to quantify the occurrences of the message. In contrast, social networks like Facebook and Twitter provide aggregated data for posts, such as the number of likes, comments, and shares. WhatsApp, however, does not offer any information that aggregates such information. To address this limitation, we propose an approach to aggregate identical messages by calculating the total occurrences and shares of each unique message.

For audio and video, after downloading the media, we calculate the checksum of each file, which enables us to determine how many times each piece of media has been shared. A checksum is a unique string of characters generated by applying a hash function. In our study, we used the MD5 algorithm to generate these checksums. This means that two identical files will produce the same hash value, allowing us to identify duplicates. By computing and comparing checksums, we can detect duplicate media files and merge identical content, ensuring a more accurate aggregation of shared messages. This approach helps us track how often the same media is shared across different groups, providing valuable insights into content dissemination.

For images, we also calculate the perceptual hash (pHash)[3]. We use pHash because, unlike a traditional checksum that relies on exact binary data, pHash analyzes the visual content of an image by capturing its structure and patterns. This allows us to identify images that are similar but may have small differences. On WhatsApp, users often share the same image with slight modifications, such as adding a logo or making minor edits. Even the process of downloading and upload an image can lead to compression, which alters the file and results in different hash values. However, pHash is more resilient to small changes.

After processing all messages and extracting the most important information, we generated a JSON file containing messages grouped by day. From the initial dataset, key information is converted into JSON files to facilitate processing and analysis. Figure 3.3

[3] https://www.phash.org/

Type	Total	Unique
Text	6,077,701 *(39.2%)*	3,274,702 *(45.3%)*
Video	3,615,603 *(23.3%)*	1,434,199 *(19.8%)*
Image	2,282,166 *(14.7%)*	1,374,543 *(19.0%)*
Sticker	1,373,889 *(8.9%)*	138,873 *(1.9%)*
Link	1,358,043 *(8.8%)*	646,043 *(8.9%)*
Audio	773,561 *(5.0%)*	579,994 *(8.0%)*
Document	26,728 *(0.2%)*	10,438 *(0.1%)*
Total Messages	**15,507,691**	**7,231,631**

Table 3.1: Summary of Message Types from January 2022 to January 2023.

illustrates the structure of a message.

3.5 Overview of Collected Messages

Considering the proposed methodology for collecting WhatsApp data, we gathered data from January 2022 to January 2023. This period encompasses key political events in Brazil, such as the presidential election, in which social networks and messaging apps were inundated with political discussions and campaigns. Furthermore, the analyzed period includes other important events, such as the intense protests marked by fraud allegations about the results and the riots on January 8th. In addition, many protests were made against the decisions of the Supreme Court of Justice and its ministers, which became a major topic in the news. By collecting data from these public groups, we can capture large-scale messaging activity and also viral content shared within this platform during the Brazilian presidential election. In total, we collected 15,507,691 million messages shared in 1,444 WhatsApp public groups monitored, as shown in Table 3.1.

Table 3.1 presents the total number of messages for each media type, along with the number of unique messages. This distinction is important because a single message can be shared multiple times. Analyzing the table, it is noteworthy that the volume of repeated messages is approximately double that of unique messages. This indicates that the same content circulates multiple times within the groups and demonstrates the dynamic nature of WhatsApp groups, where the same message can be shared repeatedly during key periods.

Considering the data collected from January 2022 to January 2023, we observed that the majority of the messages were text-based (39.2%). This is expected, as WhatsApp is a messaging app centered around text-based communication. Another significant message type consists of images (14.7%) and videos (23.3%). In our study, we focused on

Type	Total	Unique
Text	5,551,041 *(46.5%)*	3,383,562 *(45.8%)*
Video	1,876,719 *(15.7%)*	1,016,093 *(13.8%)*
Image	2,143,741 *(17.9%)*	1,381,605 *(18.7%)*
Sticker	529,717 *(4.4%)*	–
Link	1,526,572 *(12.8%)*	750,904 *(10.2%)*
Audio	295,900 *(2.5%)*	–
Document	29,161 *(0.2%)*	19,866 *(0.3%)*
Total Messages	**11,952,851**	**7,377,647**

Table 3.2: Summary of Message from March 2020 to December 2021.

political groups, as they represent a crucial source for information dissemination and are key to engaging users.

In addition to the dataset we collected, our study also incorporates the dataset from [33], which contains WhatsApp messages from March 2020 to December 2021. Table 3.2 provides a summary of this data set, which also focuses on public political groups, which are part of our collection from January 2022 to January 2023. It was originally gathered using a web interface collection method, and we further enhanced it by applying a rooting process, expanding our monitoring to additional groups, and acquiring more detailed information. In our study, this aggregated dataset served as supplementary data to examine the historical context of groups, particularly in Chapter 4, where it is used in the analysis of group hijacking attacks.

3.6 Ethical Consideration

Handling WhatsApp data requires considerations of privacy and ethical guidelines, as those messages may contain sensitive content. All data collected for this study came from public groups, where access was granted through openly shared invitation links, ensuring that no private conversations or content were monitored. User anonymity was also preserved by not storing any personally identifiable information such as phone numbers or users' real names (we only store hashes).

Moreover, in our study, we did not interact with any messages or chats. Furthermore, in the status of each account, we stated that the account is a research profile used to monitor the group. WhatsApp's terms of service allow group participants to download all content from the groups they are part of, so our data collection does not violate any of the terms. Furthermore, WhatsApp typically bans accounts that violate its terms or abuse the platform. It is important to note that during the entire data collection pe-

riod, our accounts were never banned, which indicates that our activities complied with WhatsApp's guidelines.

Studying the dynamics of political discourse and misinformation on social media also requires careful attention. Our keyword list includes a balanced number of terms for different political alignments in Brazil to ensure greater diversity of groups and avoid methodological bias. Despite this, the number of right-leaning groups on WhatsApp in Brazil is more prominent. However, our methodology accurately reflects the Brazilian political landscape on WhatsApp, which is not symmetric.

An important point is that the dataset is not released, as the potential risks outweigh the possible benefits. The data may contain sensitive information, and we do not have a complete understanding of the entire dataset (e.g., we are not fully sure what the prevalence of hateful content in this dataset is). Given the controversial nature and the potential inclusion of harmful or extremist content in our dataset, we can not guarantee that all sensitive data has been anonymized, which could also violate the General Data Protection Regulation (GDPR) law[4]. Furthermore, another reason is our concerns about the potential misuse of the dataset, including the incitement of violence and the spread of harmful ideologies. However, we believe that our work's benefits outweigh the potential harms that may arise. Our work aims to shed light on the dynamics within WhatsApp that can harm WhatsApp groups, which is important because it helps raise awareness among users about these attacks, as well as encourages platforms to improve their governance and moderation tools in an attempt to prevent or mitigate such attacks.

Our published studies on WhatsApp were approved by the ethics committees of the Max Planck Institute for Informatics (MPI), in collaboration with our research group and researchers from these institutions.

3.7 Limitation

One major challenge is the inherent difficulty in determining the representativeness of our dataset, a common issue faced by studies focusing on messaging platforms like WhatsApp. This challenge arises from the lack of a comprehensive or random sampling method to capture data from all WhatsApp groups discussing Brazilian politics. Moreover, WhatsApp primarily consists of private chat conversations, with only a small portion being public groups. As a result, our sample of groups may not fully represent the characteristics of all WhatsApp ecosystems, especially since we selected groups focused on political discussion. Despite this limitation, we believe that our data collection efforts

[4] https://gdpr-info.eu/

are extensive and enable us to demystify and characterize the main activities occurring within political groups on WhatsApp.

Another limitation is the set of keywords used to identify political public groups. New public groups emerge daily, and discussions in society evolve due to events and changing situations. Additionally, groups change, and users may lose interest in older groups or leave because they are no longer relevant. Therefore, we continuously search for and join new public groups to keep the data up to date with current trends. We can also add new keywords to the list to monitor recent events or emerging topics. Furthermore, the data collection relies on physical smartphones, and to scale the number of groups, additional smartphones are needed to access and join the groups. Since groups receive many messages per day, having a single smartphone manage all the groups can significantly slow down the process and even cause crashes, which makes collection more difficult.

Despite these limitations, the data collection presented in this work represents a large dataset that observes important recent political events in Brazil. To the best of our knowledge, this is one of the largest recent WhatsApp datasets. While it may not be fully representative, it offers valuable insights and transparency into this closed network.

Chapter 4

The Role of Mobilized Attacks in WhatsApp's Ecosystem

The rise of messaging applications has significantly transformed how people communicate and interact. The immense popularity of WhatsApp, coupled with the nature of the communication it facilitates, has created a highly convoluted and fertile environment for the propagation of misinformation campaigns. The environment created by messaging apps like WhatsApp is inherently complex. While these platforms offer a variety of tools to facilitate the rapid dissemination of messages, they also maintain a level of anonymity that conceals the authors of these messages.

The public space within the WhatsApp platform has emerged as a hub of communication and organization, enabling the seamless coordination of activists. These public groups connected hundreds of very active users dedicated to spreading information to participants and groups, creating a backbone for information propagation within WhatsApp [102]. Misinformation campaigns feed these groups of activists, who are moved by loyalty to the preferred candidate and tend to amplify the reach of the messages received, regardless of their truthfulness. At the same time, given the polarized nature of political discussions, groups of activists organize themselves into digital groups to fight with each other and to promote hostile interactions towards opponents. The public nature of these groups means that both supporters and users with opposing political views can engage in the same group. This public space also allows malicious users to infiltrate, disrupting the dynamics of these groups and enabling attacks. This no man's land created within WhatsApp by digital militias is nearly unexplored by the research community.

This Chapter presents a comprehensive study that analyzes the strategies and attacks employed by activist groups operating in public political WhatsApp groups. By examining the dynamics and tactics used by these groups, this chapter sheds light on the complex landscape of online political engagement and the role played by these activist groups in dismantling and attacking the opposing side's groups. To address this question, we explore the available data (Chapter 3), to investigate the attacks and the dynamics of WhatsApp. Thus, Section 4.1 describe the flooding attack, a commonly employed tactic to disrupt the activity of the opponent group. This attack involves overwhelming the

group with a high volume of messages. Last, Section 4.2 presents the hijacking attack, which involves the unauthorized takeover of a WhatsApp group by a malicious user, who aims to disrupt and dismantle the group. The hijacker gains control over the group, often exploiting vulnerabilities in the group's administration and then taking destructive actions, such as removing all members or spreading harmful content.

4.1 Flooding Attack

Flooding attacks are denial-of-service (DoS) attacks that aim to overwhelm a server and cause network disruption by creating network congestion [165]. Flooding attacks are not only limited to computer networks, it is also a popular type of attack and can also affect other communication channels, like SMS [67] and chat messages on online messaging platforms. On messaging platforms like WhatsApp, attackers can infiltrate a group and send a large volume of messages quickly, disrupting the group's regular operation and making it difficult for benign users to interact and chat in the WhatsApp group. Motivated by this, we aim to identify and characterize these attacks, understanding how they are carried out and analyzing their impact on targeted groups.

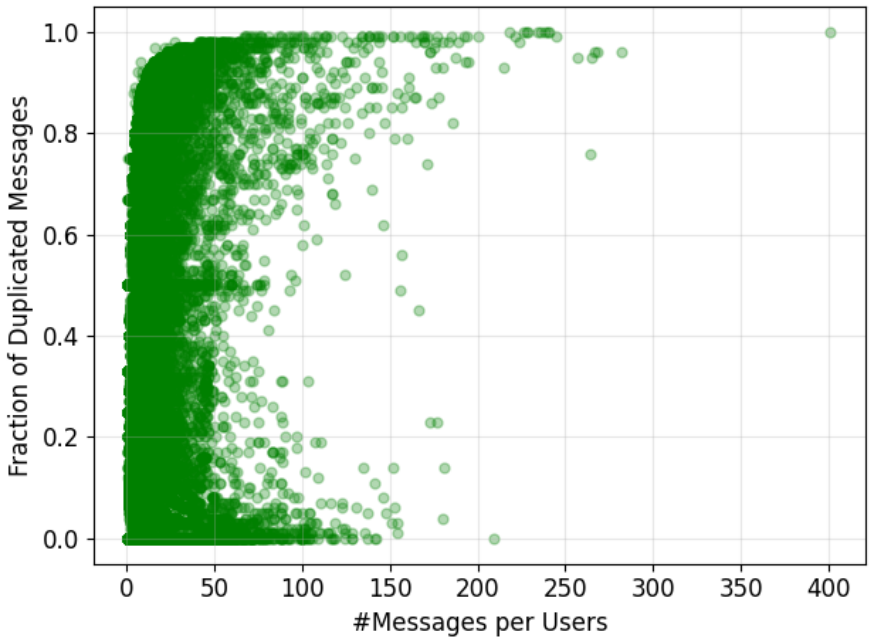
Our analysis of flooding attacks focuses on the period between July 2022 and December 2022, which includes data from 1,267 groups and 12,167,529 messages (see Chapter 3 for details on data collection). We focus on this period for some reasons. First, this is the most active period of our dataset, and second, this period in our data collection included all activity related to sticker messages, which, as we will see later, are important for detecting flooding attacks. Moreover, this period encompasses important political events, such as the Brazilian presidential elections [124] and misinformation campaigns targeting the electoral process and protests urging for military intervention [38].

4.1.1 Identifying Flooding Attacks

The flooding attack on WhatsApp group occurs when an attacker sends a large volume of messages, usually containing identical content, within a short period. To identify flooding attacks on WhatsApp groups, we devise the following methodology. First, we split the group's message-sharing activity into *sessions*, each comprising one minute of the group's activity. Then, for each session, we calculate: 1) the average number of

messages per user; 2) the fraction of duplicated messages (i.e., sharing the same text, image, audio, video, sticker).

Figure 4.1: Aggregate user-activity and fraction of duplicate messages per session.



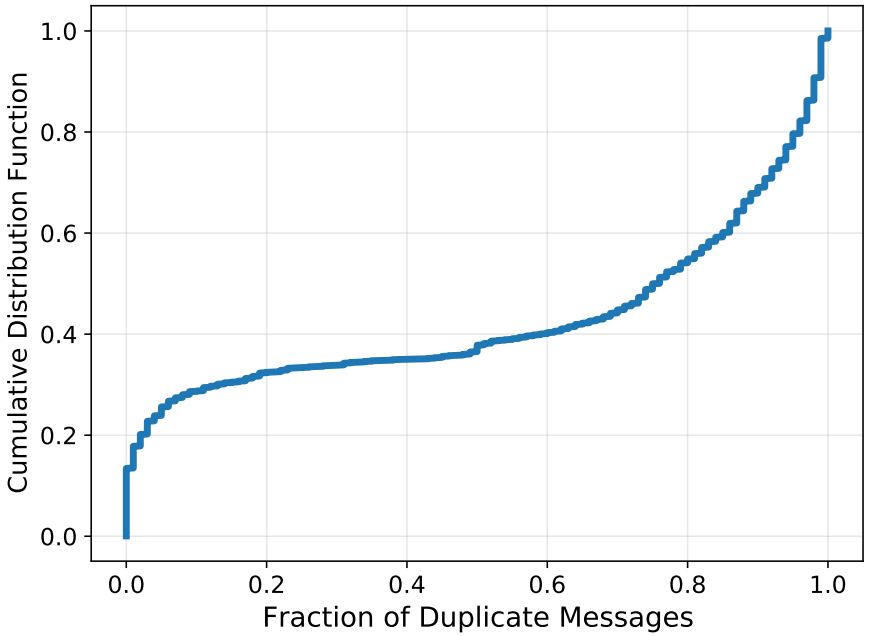
Source: The Author.

Figure 4.1 shows a scatter plot of these two metrics for each 1-minute session; we observe that the majority of the sessions have less than 60 messages per user, which is expected given that in most of the sessions, we expect to have benign conversations where many users share messages with relatively low frequency. Also, we observe that for a low average number of messages per user (less than 60), we have many sessions with a fraction of duplicated messages across the entire range of 0 and 1. On the other hand, when considering sessions with more than 60 messages per user, we observe that the fraction of duplicate messages is concentrated mainly on the limits.

This is evident by looking at Figure 4.2, which shows the Cumulative Distribution Function (CDF) of the fraction of duplicated messages for all sessions with 60 messages per user or more. We observe that 60% of sessions with 60 messages per user have at least 60% of the messages shared within the session as duplicates (i.e., users sharing messages with identical content). Based on this session-based characterization, we assume that a group is under a flooding attack when there is an average number of messages of 60 messages or more and at least 60% of all the session messages are duplicates.

Using the above-mentioned methodology, we identify 893 flooding sessions in 95 WhatsApp groups (7,04% of all active groups from July 2022 to December 2022). Then, we aggregate the flooding sessions into flooding attacks. Since we create 1-minute sessions, flooding attacks may span multiple consecutive flooding sessions. Therefore, we combine

Figure 4.2: Distribution of duplicate messages and user-activity in the 1-minute sessions.



Source: The Author.

all consecutive flooding sessions happening in the same group and treat them as part of the same flooding attack. Overall, we find 580 flooding attacks in 95 WhatsApp groups.

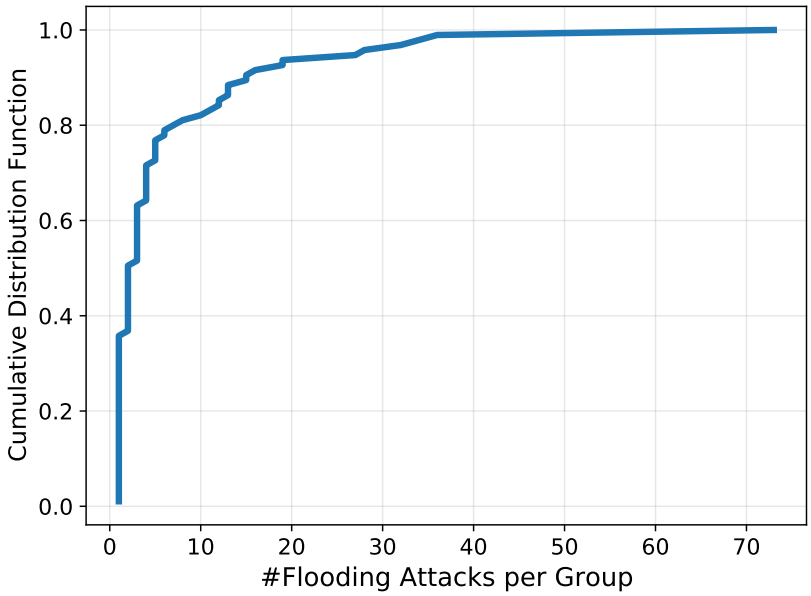
4.1.2 Characterizing Flooding Attacks

Having identified a set of flooding attacks, we aim to characterize these attacks, focusing on understanding how these attacks are executed in WhatsApp groups. We start our characterization by looking into the groups that are the recipients of the flooding attacks. Figure 4.3 shows the CDF of the number of flooding attacks received per group (Figure 4.4(a)), as well as the CDF of the number of flooding attacks per group per day (Figure 4.4(b)). Almost half of the groups experience only one flooding attack throughout our dataset.

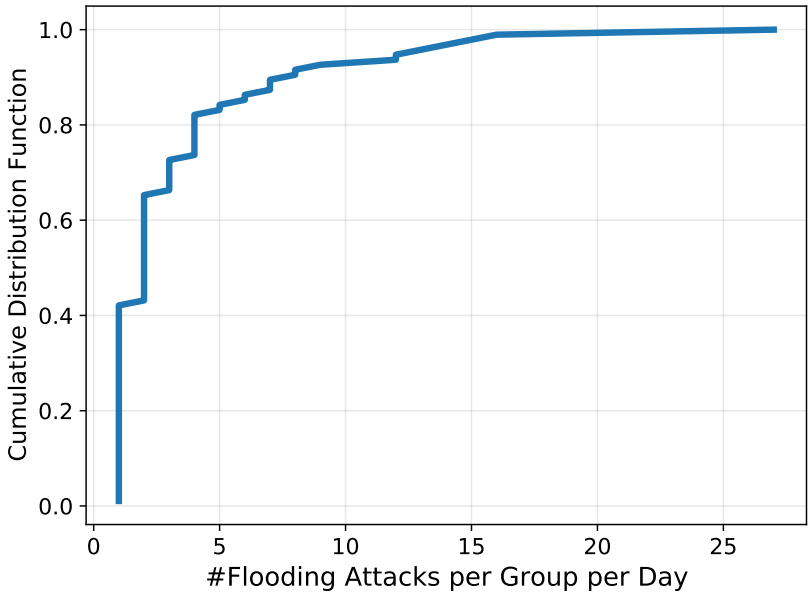
At the same time, we observe that many groups receive flooding attacks; e.g., 20% of the groups receive more than seven flooding attacks throughout our dataset (see Figure 4.4(a)). More worrying is the finding that groups receive multiple flooding attacks within the same day. We find that 57.89% of the groups receive more than one flooding attack within the same day (see Figure 4.4(b)), highlighting the prevalence and gravity of these attacks, especially in political groups.

To better illustrate this phenomenon, we present a case study of a single WhatsApp

Figure 4.3: CDF of the number of flooding attacks per group: a)for the entire period of our dataset; b) per group per day. We focus on the groups that received at least one flooding attack during our dataset.



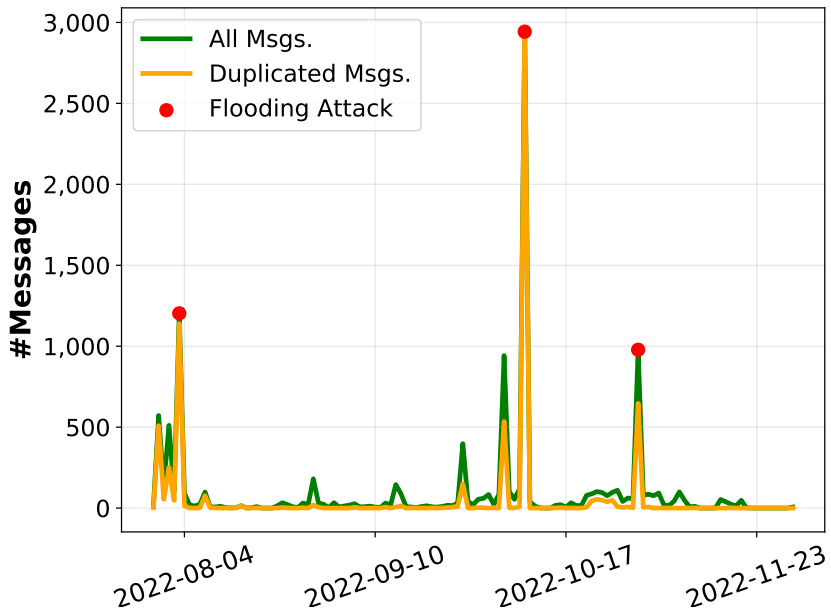
(a) Per group



(b) Per group per day

Source: The Author.

Figure 4.4: Group targeted by multiple flooding attacks.



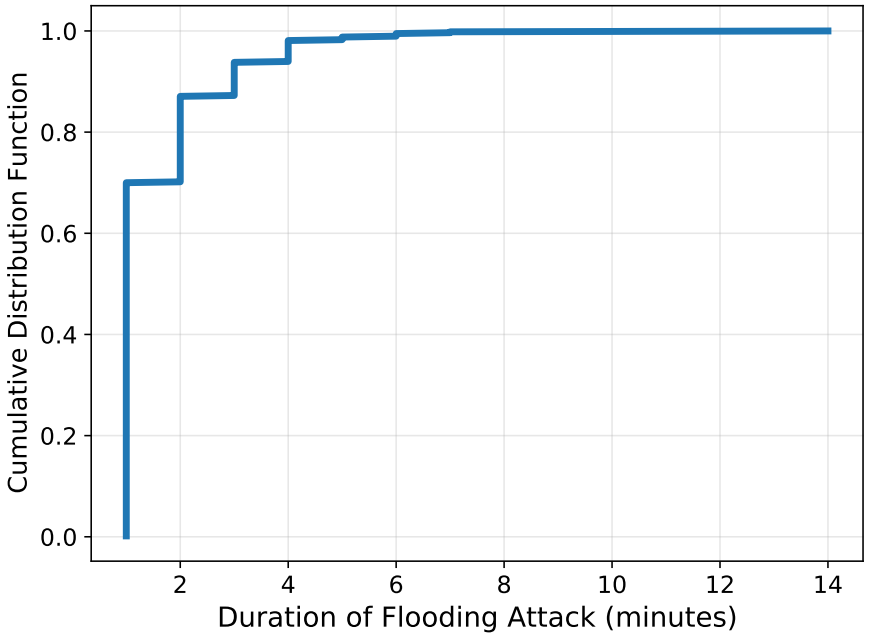
Source: The Author.

group that received multiple flooding attacks throughout our dataset in Figure 4.4. This specific group received in total four flooding attacks, with the first three increasing in intensity, as observed by the increasing number of duplicate messages shared during the attacks. Overall, the finding that WhatsApp groups are the recipients of multiple flooding attacks, and sometimes within the same day, indicates that the group's administrators can not prevent or moderate these attacks effectively. This is likely due to the absence of effective moderation tools that can assist group administrators in tackling attackers that share many duplicate messages within a short period of time [102].

Next, we look into the duration of flooding attacks. Given that flooding attacks may consist of multiple 1-minute flooding sessions, we calculate each attack's duration by summing all the consecutive 1-minute flooding sessions. Figure 4.5 shows the CDF of the duration (in minutes) of the flooding attacks observed in our dataset. We observe that, in general, flooding attacks are short-lived; 70% of the flooding attacks have a duration of up to one minute, and 98.1% of the flooding attacks have a duration of up to four minutes.

Given that flooding attacks are short-lived, we then turn our attention to looking into the type of messages that are disseminated during the flooding attacks. We expect that attackers are sending media or types of messages that can send en-masse in a short period. To shed some light on the modus operandi of the attackers, for each flooding attack, we identify the types of messages that are disseminated during the flooding attack. Figure 4.6 shows the prevalence of the flooding attacks across the various message types. Most attacks are carried out using exclusively stickers (54.13%), text (22.06%), or a

Figure 4.5: CDF of the duration of flooding attacks.



Source: The Author.

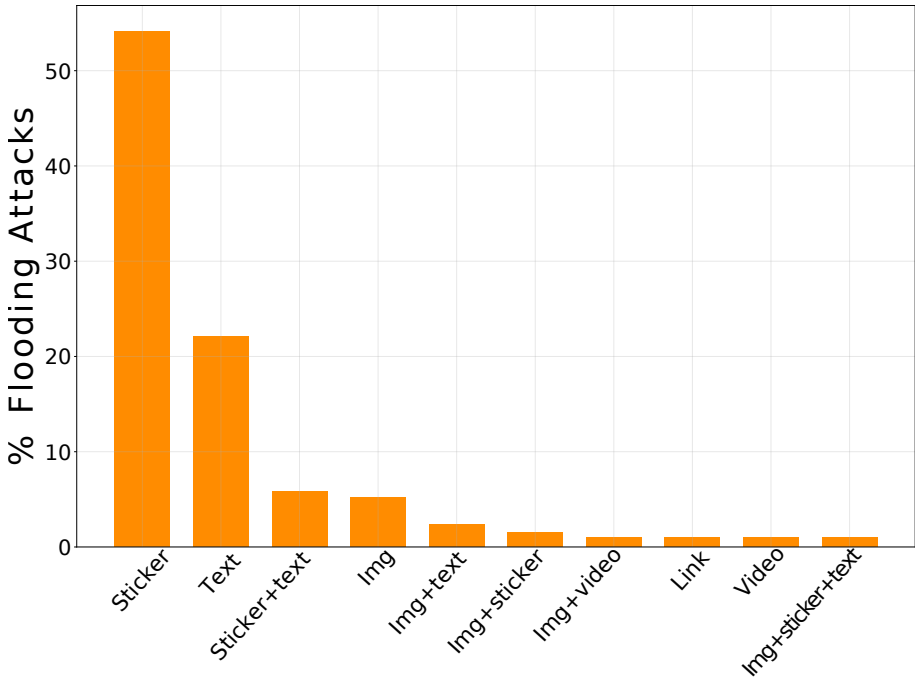
combination of both text and stickers (5.86%). Stickers are small images and can even be animated. They serve multiple purposes, like sharing emojis and memes that can be easily and quickly shared in a group. Given that stickers are also customizable, and group participants can create new stickers, attackers may create some offensive stickers and then disseminate them in the group to undertake a more severe attack.

4.1.3 Characterizing stickers used in Flooding Attacks

Thus far, we have observed that stickers are one of the most popular message types for undertaking flooding attacks. To better understand flooding attacks, here, we characterize the content of the stickers by undertaking a qualitative analysis to better understand the domains. To do this, we extracted a random sample of 100 stickers used during flooding attacks. We constructed a codebook after two researchers went through the stickers, created initial codes, discussed them together, and refined them iteratively until no more changes were made, reaching as result ten codes:

- **Political Agenda**: Promotion of political views or ridicule/criticism of the opposite political side.
- **Meme:** Non-political memes and animated content.
- **Porn:** Explicit sexual content, nudity, and pornography.

Figure 4.6: Percentage of flooding attacks for each different type of message in our dataset.



Source: The Author.

- **Abstract:** Random stickers and miscellaneous topics.
- **Disgusting:** Repulsive and disgusting content, usually involving bodily waste.
- **Religious Attack:** Employing religious symbols and irony to mock or satirize religion.
- **Violence:** Content that promotes violence.
- **LGBTQ+ Attack:** Content attacking LGBTQ+ people.
- **Antisemitism:** Promoting nazism or attacking Jewish.
- **Racism:** Content promoting racist views.

Having constructed the codebook, two researchers independently coded another sample of 100 messages; we find a Cohen's Kappa coefficient of 0.936, indicating high agreement between them, hence the rest of the stickers are coded from a single annotator. Overall, we coded all 1,667 stickers used in flooding attacks. We find that almost 60% have a Political Agenda, which shows that the attacks also aim to provoke the opposite side and sometimes promote their ideologies. Memes (17.87%) are frequently employed to inundate the group. Abstract (8.62%) stickers are readily accessible, with some being standard stickers commonly found on many phones, indicating that users choose random stickers to inundate the group.

More worrying is the fact that we find a substantial percentage of stickers containing harmful content like Porn (12.34%), Violence (1.12%), and Disgusting (2.79%), including repulsive content and offensive imagery. The attacker's goal extends beyond merely targeting the group; it also involves creating discomfort by disseminating offensive and repugnant content. Other instances of hate content include Religious Attacks

(1.18%), LGBTQ+ Attacks (0.87%), Antisemitism (0.81%), and Racism (0.50%). The substantial degree of harmful stickers used in flooding attacks highlights the gravity and potential impact of such attacks on WhatsApp users.

4.1.4 Characterizing text used in Flooding Attacks

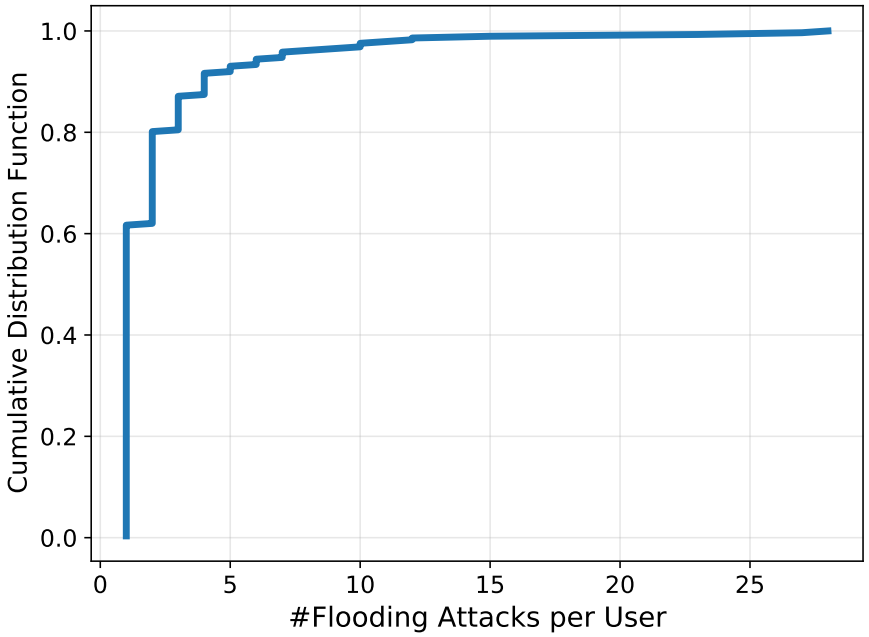
Text messages are also popular for flooding groups; 31.3% of flooding attacks used text messages. Here, we aim to characterize the text content shared during flooding attacks. To do this, we categorize 20% of all flooding text messages. Initially, two human evaluators examined 30% of these messages and established the five codes:

- **Overload Attack Message**: A long message with many characters designed to slow down the phone.
- **Political**: Messages strengthen their own political stance.
- **Meaningless**: Laughter or random characters with no specific meaning.
- **Accusation/Attack**: Messages to provoke or incite political reactions.
- **Word/Phrase**: Words or phrases lacking a particular discernible purpose.

Subsequently, all text messages in the sample were labeled. Within this set of messages, 54.88% were identified as Overload Attack. This shows the malicious intent of the attacker, who not only floods the group but also attempts to slow down users' phones. Furthermore, an additional 25.61% Meaningless messages, characterized by random characters seemingly typed on the keyboard without any discernible meaning. Another 14.64% contained Political and Accusations/Attacks designed to provoke the opposing political group. Finally, 4.88% consisted of random Words/Phrases lacking a specific meaning.

To conclude our characterization of the flooding attacks, we look into the users who participate in flooding attacks (i.e., attackers). To identify attackers, we extract the most active users in each flooding attack and we treat the user as an attacker if the number of messages they shared during the attack period is 20% or more of the entire session activity. Figure 4.7 shows the CDF of the number of flooding attacks per user, as well as how many users are participating in the same flooding attack. We find that 38.3% of the users that participated in flooding attacks participated in more than one attack throughout our dataset (see Figure 4.7). Also, we find that most of the flooding attacks are executed by a single attacker (90%, see Figure 4.8).

Figure 4.7: CDF with flooding attacks per user.



Source: The Author.

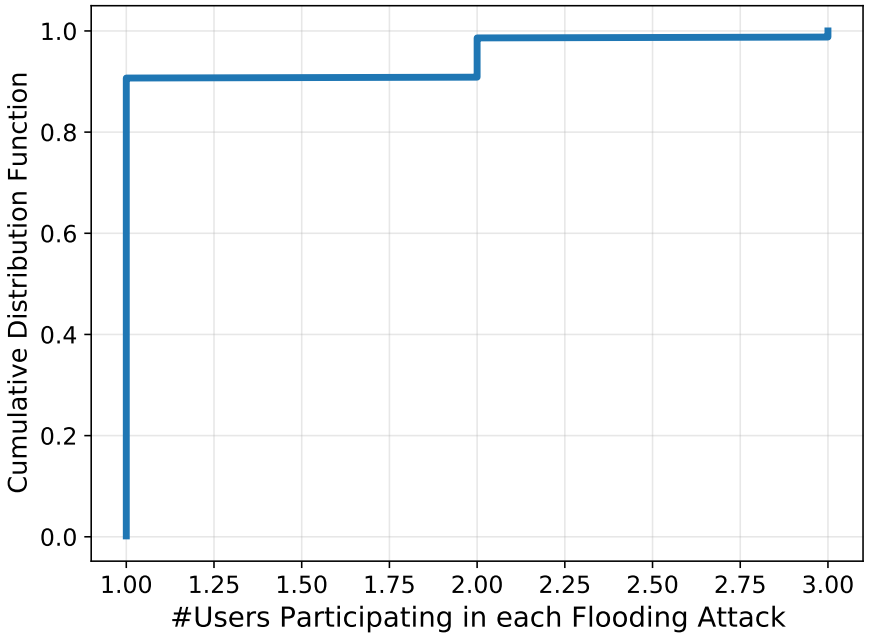
4.1.5 Impact of Flooding Attacks

Having characterized flooding attacks, here, we aim to analyze the impact of these attacks on WhatsApp groups. To achieve this, we compare the activity within the group before and after each flooding attack. Given that we already showed that flooding attacks are short-lived, we focus on 24 hours before and after the attack. Then, we calculate the hourly number of messages and active users and present the results in Figure 4.9 and in Figure 4.10. We focus here on cases where we only have one flooding attack within 48 hours (corresponds to 21.03% of all flooding attacks), as subsequent flooding attacks will affect the WhatsApp group activity.

We observe a substantial increase in both metrics (number of messages and active users) during the flooding attack. This result is expected for the number of messages, given that the attack itself is based on the creation of a large number of messages. On the other hand, the increase in the number of active users is likely due to benign users enquiring about the attack. After the attack, we observe that the group is restored to its regular activity three hours following the flooding attack; on aggregate, we have a similar number of messages and active users before and after the flooding attacks. Overall, these results highlight that flooding attacks can potentially disrupt the groups' activity, however, their impact appears limited to only a few hours.

To conclude our analysis of the impact of the flooding attacks, we perform a small-

Figure 4.8: CDF with users per flooding attacks.



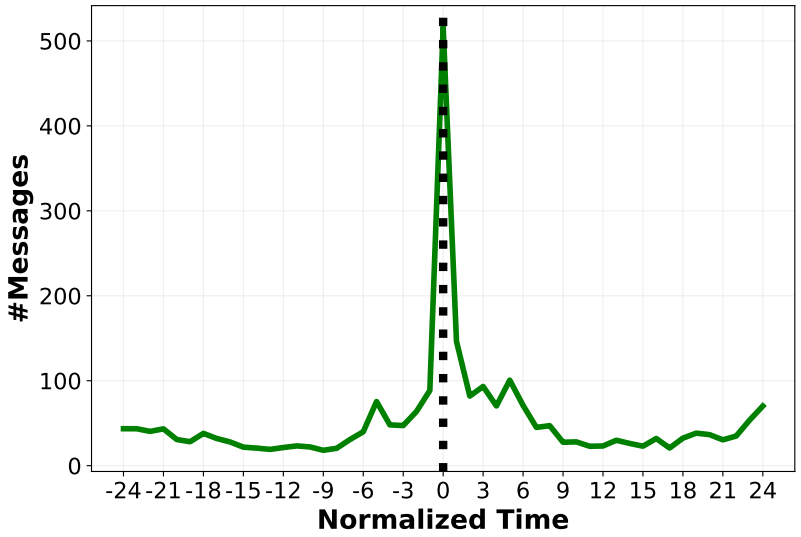
Source: The Author.

scale qualitative analysis of text messages sent by benign users to shed some light on the impact on WhatsApp users. In this direction, we labeled 20% of all text messages sent by benign during the flooding attack, corresponding to 243 messages. Initially, 30% of messages were labeled to find these five codes:

- **Complaining/Cursing**: Users express dissatisfaction or use profanity to describe flooding attacks.
- **Requesting Moderation**: Users request the administrator to take action and eliminate the attacks.
- **News**: Political news or new media forwarded.
- **Interaction**: Users engage with one another.
- **Miscellaneous**: Laughing, meaningless content or message that we could not identify.

Next, all other messages were labeled by two human evaluators, reaching a kappa coefficient agreement of 0.70 [26]. Among all the messages, 31.28% specifically pertain to Complaining/Cursing about the ongoing attack, while 13.17% of messages were out of a lack of Moderation and requested the administrator's intervention to remove the attacker. For instance, some examples we observed during flooding attacks from benign users are: *"Where is the admin?"*, *"Hey admin ban the attacker"*, *"What is it?"*, *"my cell phone is crashing"*(translated messages from Portuguese). Furthermore, 25.93% of the messages involve users interacting with each other or responding, while 16.05% consist of News content shared during the attack. Lastly, 13.58% constitute Miscellaneous, which could not be clearly identified or categorized.

Figure 4.9: Number of messages before and after flooding attacks. We normalize the time and we focus on 24 hours before and after each attack (time 0 corresponds to the flooding attack).

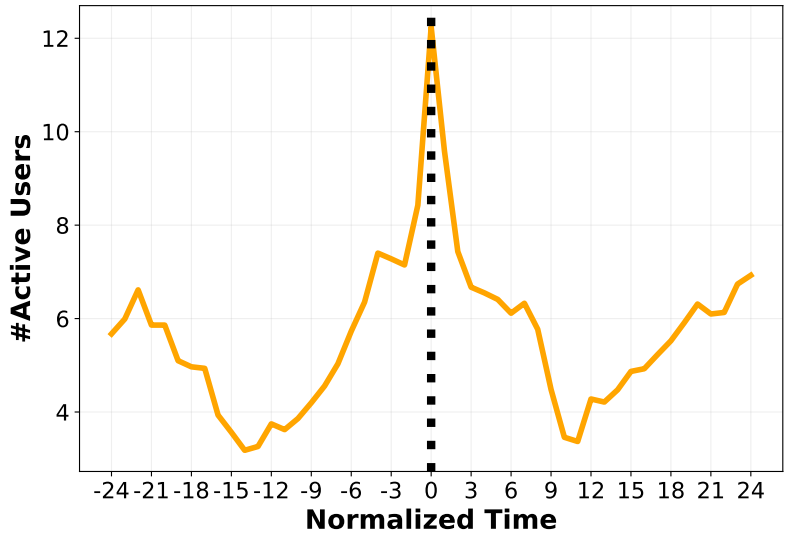


Source: The Author.

4.1.6 Flooding takeaways

- Flooding attacks are not a rare phenomenon on WhatsApp when considering Brazilian groups related to politics. We find that 7.04% of all monitored WhatsApp groups experienced at least one flooding attack throughout our dataset.
- Flooding attacks are short-lived (70% of the flooding attacks have a duration of up to one minute), and they are usually undertaken by a lone wolf (i.e., 90% of all flooding attacks consist of a single attacker).
- We find that WhatsApp groups are the recipients of multiple flooding attacks (even within the same day), which indicates that the moderators or group administrators cannot proactively prevent flooding attacks.
- Many flooding attacks are made using the large dissemination of stickers; 62.57% of all flooding attacks in our dataset use stickers. In addition, based on our manual annotations, we find flooding attacks that use offensive stickers including pornography, violence, and provocative messages.
- By analyzing the impact of flooding attacks, we find that the irregularities in the group activity due to the attacks last for only a few hours (usually three hours), which indicates that both the attack and its impact are short-lived.

Figure 4.10: Number of active users before and after flooding attacks. We normalize the time and we focus on 24 hours before and after each attack (time 0 corresponds to the flooding attack).



Source: The Author.

4.2 Group Hijacking Attack

In the previous section, we investigated flooding attacks on WhatsApp groups and observed that WhatsApp groups tend to return to regular operation/activity after the flooding attacks. Here, we investigate a more severe attack, the Group Hijacking Attack, where an attacker aims to obtain complete control of the group and potentially make drastic or catastrophic changes. The attack is initiated when an attacker obtains administrator rights in the group either via unauthorized means (e.g., by compromising the account of an administrator or the group creator) or by enticing other administrators to promote the attacker to an administrator, or by identifying groups where the creator left the group. Having obtained administrator rights, an attacker can make important changes in the group, such as removing group participants, repurposing the group by changing group metadata, or even archiving/deleting/privatizing the group.

The group hijacking attack is analogous to attacks aiming to compromise accounts on social media platforms [37] to either repurpose it [39], steal personal information, or share misleading information. Here, we focus on understanding and characterizing this phenomenon through the lens of political groups from Brazil on WhatsApp. In political groups, hijacking attacks can be used to disrupt the discussions of supporters from the other party or attack them by sharing provoking content within their group. Overall, there is a pressing need to understand this phenomenon, as it has the potential to increase

online political polarization in the WhatsApp ecosystem. Our analysis of hijacking attacks focuses on the entire collected dataset, including messages collected between March 2020 and December 2022, with 189k active users.

To identify potential group hijacking attacks, we focus on groups that had at least one change in the group's name throughout the period of our dataset. We find that 563 groups (34.28% of all groups) had at least one name change; not all name changes pertain to attacks. To identify potential attacks, we then manually annotate all the 563 groups and their respective name changes to identify whether the name change is suspicious (e.g., the group's name before and after the change substantially differ semantically) by two evaluators with a 0.7 kappa coefficient. We find a total of 33 groups with substantial semantic differences in the two group names, which corresponds to 5.86% of all groups with at least one name change in our dataset.

4.2.1 Characterizing groups with name changes.

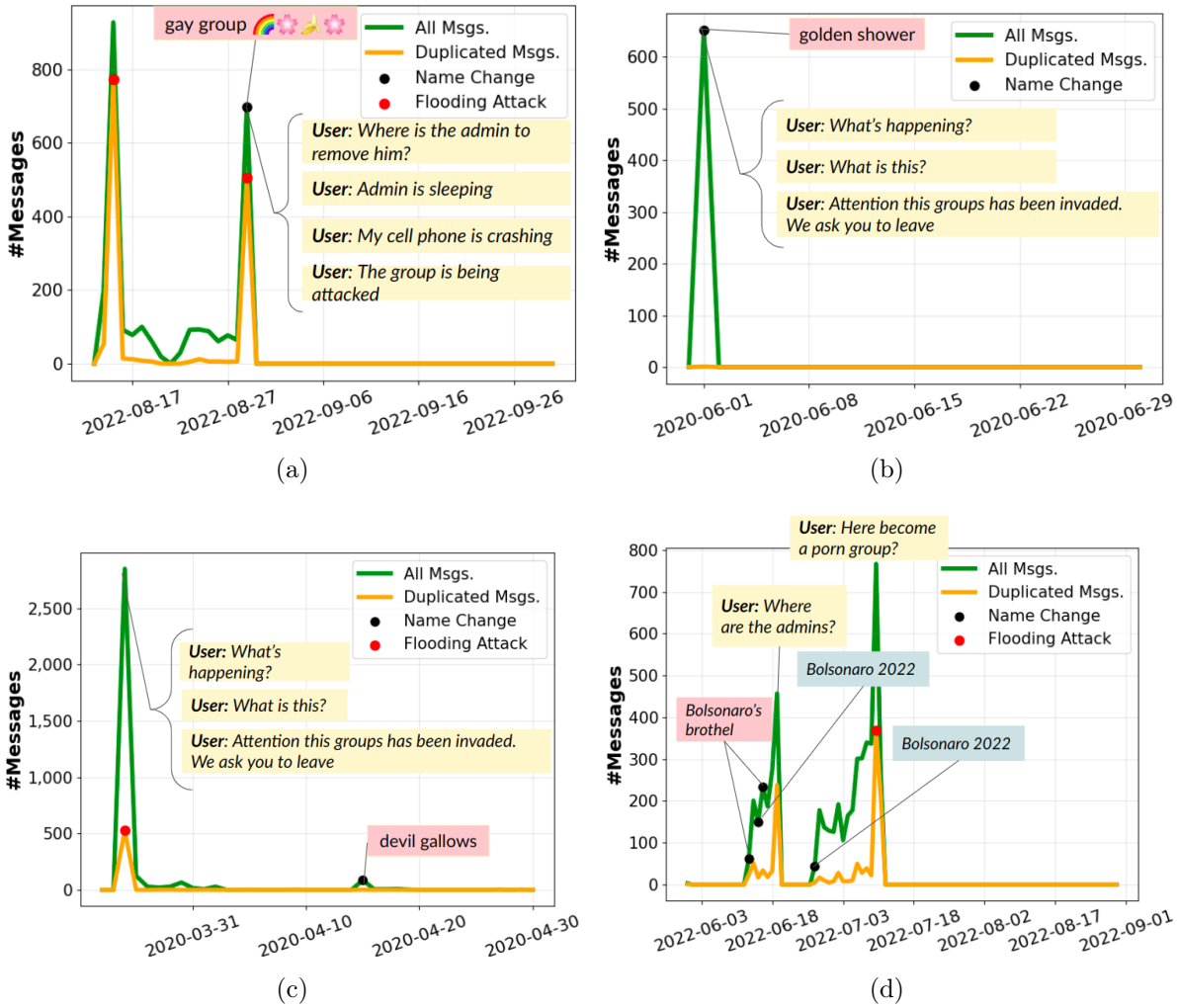
Based on our manual annotations, we categorize these 33 groups into four high-level categories (see Table 4.1).

- **Conflicting:** These are groups that, after the group name change, a contradiction emerges in the political ideology when comparing the period before and after the change in the group's name. We find four cases of groups in our dataset with conflicting political leanings. For instance, a group named "Bolsonaro reigns" was renamed to "Left Reigns," indicating the substantial shift in the group's ideology, likely because of a group hijacking attack.
- **Offensive:** Groups where the group's name became more offensive or toxic after the name change. We find in total five such cases of groups in our dataset. For instance, we find a group that was changed from "Bolsonaro 2022" to "Gay Group" and a group changed from "Patriots and the Captain" to "bordel bozo" (Bolsonaro's brothel).
- **Explicit:** These are groups where the name change indicates that there was an attack on the group. We find four such cases; some examples include a group initially named "Alliance for Brazil" and then "Hacked" and a group renamed "Archived by apocalipse."
- **Context Switch:** Refers to groups where the name change indicates a substantial shift in the group's context and topic of discussion. We found 15 groups with context switching. For instance, a group called "We with Bolsonaro" was renamed to "Grandma's recipe," a topic that has nothing to do with political discussions.

Categories	Original Name	Changed Name
Conflicting	*Bolsonaro's slogan*	*Lula 2022*
	Evangelic & Lula	*100% Bolsonaro*
	Brazil Patriot	*Antifa Action*
	Bolsonaro the myth	*Arthur King*
	Bolsonaro Reigns	*Left Reigns*
	Bolsonaro's Slogan	*Brazil is Lula*
	Lula vs Bolsonaro the Myth	*Lula big vs Bolsonaro*
	*Haddah is a sh***	*Antifascist*
Offensive	*Bolsonaro 2022*	*Gay Group*
	Alliance for Brazil	*Golden Shower*
	Captain gallows	*Devil Gallows*
	Just patriots	*Bozo the shit*
	22	*Prostitution Group*
	Bolsonaro 2022	*Bolsonaro's Brothel*
Explicit	*Entourage PT*	*157 by: lagxzada*
	Alliance for Brazil	*Hacked*
	Anything goes!	*Archived by Apocalypse*
	Itu antifascism	**bot.py**
Context Switch	*We with Bolsonaro*	*Grandma's recipe*
	Antifascist Alliance	*Banana Cake*
	Beloved Brazil	*Birthday Uncle Dudu*
	Strategic Right	*Cake Recipes*
	Brazil-CE	*Yoga Group*
	Bolsonaro President	*Play the Siri*
	Brazilian patriots generation	*New/used online business*
	Bolsonaro's power 20 years	*Zumbussiness your future*
	Civil Resistance!	*Coconut Water*
	Anti Communists	*Groups of Friends*
	Bolsonaro News	*Max iptv and netflix*
	Politics Revealed Youtube	*Free Consultancy*
	300% Bolsonaro patriots	*Happy Family*
	United Right	*Grandma's Recipe*

Table 4.1: Groups with suspicious name changes (translated names from Portuguese).

Figure 4.11: Examples of hijacking attacks (offensive name changes). Figures (a), (b), (c), and (d) show examples of offensive name changes.



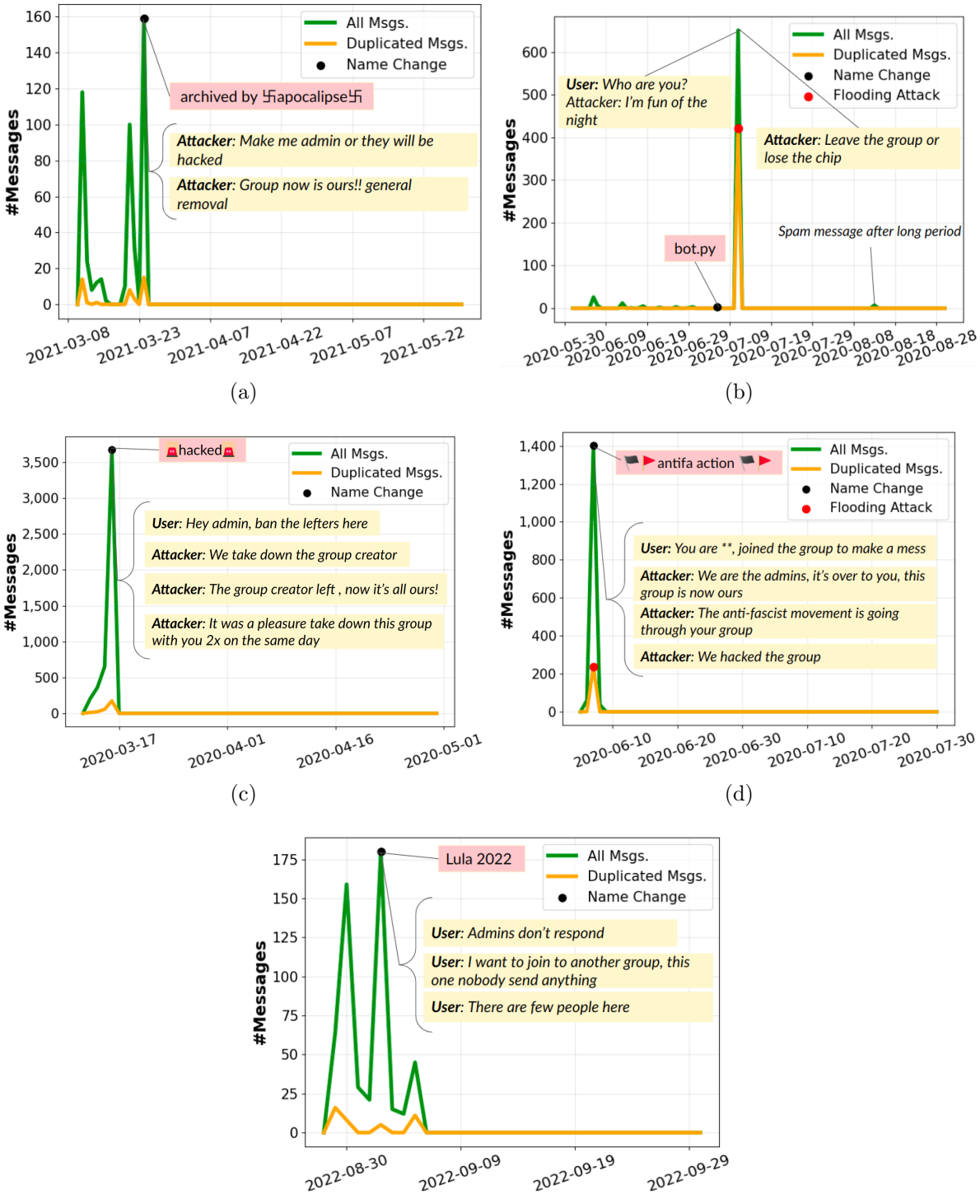
Source: The Author.

4.2.2 Identifying and annotating Hijacking Attacks.

Just because some WhatsApp groups have suspicious changes in their metadata does not necessarily mean they are the victims of hijacking attacks. Therefore, to detect hijacking attacks, it is paramount to analyze and understand what happened in the WhatsApp groups after the name changes and what messages were shared (if any) after the name change. To do this, we performed a manual annotation on the 33 WhatsApp groups that had suspicious name changes based on our previous annotations. In particular, for each group, we plot and evaluate the message activity (i.e., number of messages shared per day, before and after the name change) and manually read messages before and after the name change.

Our analysis yields several interesting observations. First, we find that 15 out

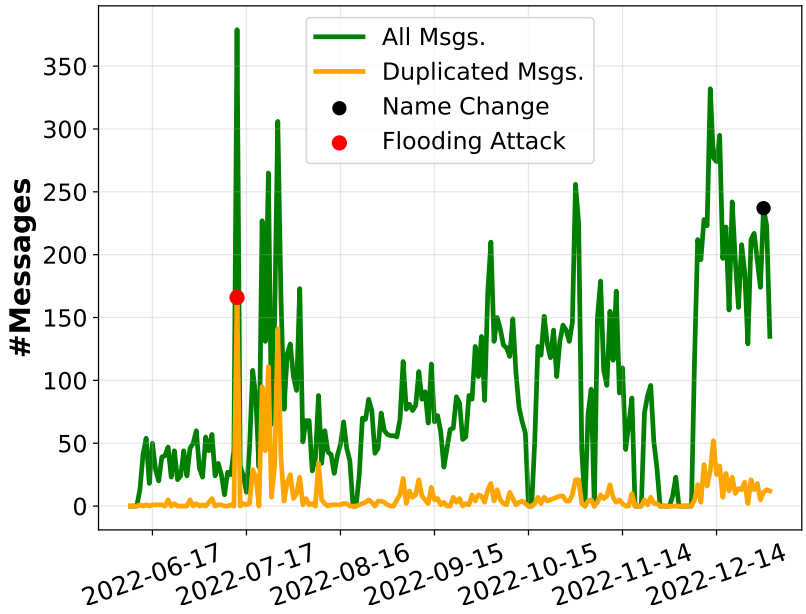
Figure 4.12: Figures (a), (b), and (c) show examples of explicit name changes, and (d) and (e) show examples of conflicting name changes.



(a) (b) (c) (d) (e)

Source: The Author.

Figure 4.13: Switch Context Example.



Source: The Author.

of the 33 groups that had suspicious name changes pertain to a group hijacking attack. For 15 of the groups, we observe that after the hijacking attack, the activity (in terms of messages shared) of the group becomes zero and we find messages that indicate that indeed an attack occurred. For instance, Fig. 4.13(c) shows an example of a group that received a hijacking attack, and shortly thereafter the group is destroyed and has no further activity.

In addition, the attacker provokes the participants who complain and say that the group was overthrown, in addition to celebrating the success of the action (see yellow box in Fig. 4.13(c)). The rest of the groups (all part of the Context Switch category), we observe that are not the victims of group hijacking attacks, but rather, the group had a name change in an attempt to obfuscate the political nature of the group. Below, we present some examples of hijacking attacks and examples of context switching.

4.2.3 Characterizing Hijacking Attacks.

Figure 4.11 and Figure 4.12 shows 9 out of the 15 groups that were the recipients of hijacking attacks (others omitted due to space). For these suspicious names, we have read all messages on the days before and after the group name change. The messages sent on the attack day help us understand what happens in the group. Looking at Fig-

ure 4.12(c), 4.13(c), 4.13(a) and 4.13(d), we realize that the attacker sometimes interacts with the participants, threatening them *"make me admin or they will be hacked"*, sending messages that the group was taken down *"it's over for you, this group is now ours"*, or even celebrating successful group destruction, *"It was a pleasure to take down this group with you 2x on the same day"*. In other cases, the attacker does not interact, but messages from participants indicate that an attack is taking place. Figure 4.12(a) shows an example where some users ask the admin's support to moderate and remove the attacker because the group is under attack. In Figure 4.12(b), the last message was sent by a user informing the group was attacked and the participants needed to leave.

Upon careful observation, we have identified a pattern in hijack attacks: most groups had an attack shortly after we joined the group, typically within a maximum of 10-20 days from its initial creation, 11 of 15 hijacking groups. Moreover, 9 out of the 15 hijacked groups ceased their activities within a maximum of 5 days. In Figure 4.13(c), we see an example of a group that suffered an attack on the same day of its creation, and the group was destroyed. This pattern suggests that attackers are specifically targeting recently formed groups. The public group is shared on social networks, and many users join, even malicious users, who take advantage of the recently created group to ask for help with group moderation and gain admin privileges. Each group contains at least one admin responsible for moderating the group, but others can also be added to help organize the group. The attacker can gain the confidence of the moderator by assuming the role of an admin within the group, once a malicious user obtains admin privileges, the group becomes vulnerable to destruction. In some cases, the group has vulnerabilities and, by default, allows users to join with administrator privileges.

Moreover, the attacker can intimidate group members and foster a hostile environment. In Figure 4.12(c), we can observe how the attackers orchestrate a flooding attack toward the admin's private chat. This can coerce the admin into either leaving the group or adding the attackers as additional admins. In Figure 4.13(c), the attackers also flooded the group and sent a message stating that the group's original creator had been removed and that the group now belonged to the attackers. In two other groups, the attacks did not occur immediately after their creation, but within days of joining the groups. This suggests a scenario where the group admin re-shared the invite link to attract new users, and attackers discovered a new group to infiltrate.

Out of the 15 groups that were targeted in the attacks, one group managed to avoid complete destruction because the admin promptly took action by renaming the group and removing the attacker's presence. This case shows that quick administrator action can help mitigate the damage caused by an attack. In practice, we realized that the attacked groups did not have an active and engaged administrator, which facilitates the action of the attackers. In Figure 4.12(d) and 4.13(e), we show how users are complaining about admin action: *"where are the admins?"*. In Figure 4.12(a), the user goes so far as to say

that the administrator is sleeping.

Finally, among the hijacked groups, we found that 4 also had a flooding attack on the day the group was renamed. This suggests a strategy employed by the attackers: hijacking and flooding attacks are used to gain control and disrupt the group. By flooding the group, the attackers create chaos and confusion, facilitating their hijacking actions. Flooding attacks are not just random messages; sometimes they are selected to evoke fear or intimidation among the group's users. In Figure 4.12(d), the attackers flood the group with pornographic messages and stickers.

4.2.4 Characterizing context switching.

Looking for cases that had significant name changes but were not identified as an attack, we find context-switching groups. We noticed that 11 out of 15 cases occurred on specific days, coinciding with either the second round of the Brazilian presidential election or the final days of 2022 when Bolsonaro exited the Brazilian government. During that period, the Superior Electoral Court (TSE) issued numerous court orders to shut down several WhatsApp and Telegram groups used to coordinate protests alleging election fraud [100], which resulted in the invasion of the Brazilian Congress, Supreme Court, and presidential offices [111].

These groups argue that there may be prosecution and therefore they need to camouflage to avoid censorship by the next government. Upon examining the activity within these groups, it becomes evident that despite the name change, they sustained a consistent level of message exchange and active user participation (see Figure 4.13). The change in group names serves as camouflage, offering participants a sense of confidence and security to express their thoughts and opinions freely. By adopting new names, these groups seek to preserve a level of anonymity and protect themselves from potential sanctions [100].

4.2.5 Hijacking takeaways

- The hijacking attack targets recently formed groups, 11 out of the 15 hijacked groups being compromised shortly after we joined the group, either within a few days or 10-20 days of the group's creation.

- The hijacking attack goes further than flooding attacks, taking control of the group and disrupting the overall interaction. In 9 of 15 hijacking attacks, the group activity stops within a maximum of 5 days.
- 11 of 15 of the groups classified as context switch did this on presidential election day or the last day of Bolsonaro's presidency. These groups are from Bolsonaro supporters, and by adopting new names, aim to maintain a certain level of anonymity and shield themselves from possible sanctions [100].
- Among the hijacked groups, 4 also had a flooding attack on the same day as the group renaming, indicating a connected strategy to disrupt and gain control over groups.

4.3 Summary

In this Chapter, we explored two kinds of attacks that can disrupt the activity of WhatsApp groups, particularly flooding attacks that aim to disseminate many messages within a short period and hijacking attacks that aim to take control of the group and drastically change its purpose. We collected a large-scale dataset of 1.6K WhatsApp groups related to Brazilian politics between March 2020 and December 2022, including 19 million messages. Then, we propose a methodology to identify and characterize flooding attacks and investigate hijacking attacks by focusing on WhatsApp groups.

The analysis shows that flooding attacks are not rare when considering political groups in Brazil. It is likely a way for people from one party to attack people from another party, aiming to disrupt their conversations. We find that flooding attacks are usually short-lived, and most flooding attacks are made by one attacker. Also, we find that WhatsApp groups are the recipients of multiple flooding attacks, even within the same day, which likely highlights the lack of effective tools that assist the group moderators and administrators who aim to maintain the group's harmony. Regarding hijacking attacks, we find many such attacks in our dataset, and we find that in most cases, the attacker's goal is to close or remove the group. Finally, we find that WhatsApp groups can be the recipients of both flooding and hijacking; the attackers first flood the group and then hijack the group entirely.

Overall, the study is a significant leap towards demystifying the dark side of WhatsApp groups, particularly hostile intergroup interactions across the political spectrum. This study highlights the prevalence and gravity of these attacks, identifying that they are not rare in Brazilian political WhatsApp groups. Additionally, our qualitative analysis highlights that a significant portion of these attacks contains harmful content, such

as hateful or offensive stickers, as well as overly long messages to disrupt WhatsApp's regular operation on users' phones. Taken together, these findings show that these attacks are an important problem, and our work has the potential to raise awareness about these attacks among both WhatsApp users and operators of messaging platforms. For WhatsApp users, raising awareness of these attacks is important as it allows them to be more prepared when the attack is underway and try to protect themselves. For messaging platforms, our work and findings can be used to raise awareness about how these attacks are performed on their platforms, which is vital for designing effective moderation tools.

Chapter 5

Understanding the Use and Abuse of Stickers

In the Chapter 4, we unveiled how malicious users exploit stickers as an attack form in public WhatsApp groups, particularly during flooding attacks. Stickers have emerged not only as a popular media type for interaction but also as a potent tool in the dissemination of harmful, offensive, and politically provocative content. Our findings revealed that more than half of the flooding attacks in our dataset employed stickers, many of which conveyed political propaganda, hate speech, or explicit imagery intended to provoke, harass, or destabilize target groups.

Building on these findings, this chapter presents a deeper and systematic characterization of sticker usage on WhatsApp. Rather than focusing solely on attack dynamics, we now turn to a broader analysis that examines stickers as a distinct form of media, exploring their structural, visual, and behavioral patterns. Our goal is to understand how stickers differ from other media formats, how they are created, shared, and collected, and how their content and visual features reflect broader sociopolitical dynamics within the messaging platform.

WhatsApp introduced stickers on the platform referring to them as something to "*help you share your feelings in a way that you cannot always express with words*".[1] The platform enables users to create custom stickers on any subject or occasion, allowing them to be quickly integrated into various contexts. This flexibility has made stickers particularly prominent in political discussions, which is a widely popular topic on WhatsApp [126]. This becomes even more relevant considering that WhatsApp has frequently been associated with the rapid spread of misinformation [11] and has played a central role in political disinformation campaigns [16, 103].

In Brazil, the popularity of stickers has increased considerably. Brazilians hold the record for the highest number of stickers sent [21], transcending their role as a mere form of humor to become a key element of political strategy. The 2022 Brazilian presidential election marked the first major political event in which stickers were massively deployed, with political actors using them to disseminate campaign messages from a visual and

[1] https://blog.whatsapp.com/introducing-stickers

emotional perspective [28]. Their influence has grown so significantly that the Brazilian Electoral Court has released an official sticker pack to aid in fact-checking news [143]. In this direction, stickers have even been weaponized in political activism campaigns, with supporters of one political party using flooding and hijack attacks to spam and undermine their opponents within WhatsApp groups during elections, as shown in Chapter 4. This open and permissive environment, in which users are free to create and share custom stickers, has also led to serious abuses. Some users have produced offensive and hateful stickers that threaten group members, including child sexual abuse material [43] and Nazi propaganda [136], creating an unsafe environment for users. As stickers have become increasingly popular among users, we have observed a rise in their misuse on WhatsApp, ranging from misinformation campaigns to hateful harassment content. However, we still lack a comprehensive analysis of their usage on a broader scale within the messaging platform ecosystem, due to the closed, private, and encrypted architecture of instant messaging platforms.

To address this gap, this chapter aims to provide a comprehensive perspective on the dual nature of stickers: both as tools for enriching digital expression and as vectors of abuse in politically charged environments. Through this analysis, we seek to answer the following questions: How are stickers shared in public WhatsApp groups? How is political content visually encoded in these artifacts? And how do users exploit stickers to propagate offensive or harmful content?. To this end, we focus on messages shared during the 2022 Brazilian presidential election period, identifying over 650,000 sticker messages encompassing 57,031 unique stickers.

We begin by analyzing sticker usage patterns and volume across our dataset, identifying peaks in activity during major political events such as the 2022 presidential election (Section 5.1). We then present a visual content analysis by clustering stickers based on perceptual similarity and dominant color, uncovering visual trends associated with political campaigns and expressive conventions (Section 5.2). Next, we investigate the political attacks made with Stickers (Section 5.3). Finally, we explore how these stickers encode ideological messages and are strategically propagating offensive or harmful content (Section 5.4).

5.1 Stickers as a Distinctive Form of Media

Multimedia messages are very popular on WhatsApp [66]. While text messages remain the majority, users often share different kinds of media using images, videos, audio, and stickers. Although images and stickers are both visual media, some key changes

distinguish them. Stickers may display animated images in chats. They are smaller than actual images and are stored in a different format by the app. Unlike regular images, stickers are usually displayed directly within the text, similar to emojis.

Stickers are usually described as emoticons in the form of colored images [24], as both can be used as visual reaction messages or determine the feeling of the interlocutor during the conversation, but differ from in-line emojis in diversity, complexity, and usage [86], providing a richer communication [154]. Also, they differ from simple static images as they can be short animated movies. Stickers are also closely related to memes [155], which often present a combination of a picture and a statement, typically with sarcastic or humorous intention [32]. Usually, there is a template-based image that people modify, edit, and publish their version while keeping some key aspects so that the meme template is still recognizable.

Both memes and stickers can be created based on personal experiences, but inspiration can be obtained from popular cultural products such as television shows and video games [55]. Additionally, users publish memes on the Web to create an incentive for others to share by replicating the original [87]. The origin of these stickers dates back to 2011, when LINE, a popular messaging app in Japan, mixed cartoons and "emojis" [130]. This tool allowed users to express socio-cultural differences in a more specific way, inspiring other applications to adopt stickers as well. In 2013, Facebook added stickers to its platform, but still with a limited set of pre-created images. This media was adopted by WhatsApp only in 2018.

Although they share similarities, stickers also diverge from GIFs in many ways. WhatsApp distinguishes between reaction GIFs and stickers by incorporating both separately in its interface, highlighting important differences in their nature and use. Reaction GIFs, a popular form of internet communication [9], are integrated into WhatsApp via Tenor,[2] an online GIF library owned by Google. When a user searches for a GIF in WhatsApp, they are searching in this database. The selected GIF from the cloud is then embedded in the message and sent. Stickers, on the other hand, are not hosted externally. Instead, they are static or animated WebP image files stored in the user's private sticker collection on WhatsApp and also in their devices. As a result, users can prohibit the download of any kind of media file except for stickers [81]. Users can create their own stickers using WhatsApp's built-in tool or third-party apps, or they can save stickers received in chats to their collections. Because of that, stickers on WhatsApp offer a much greater degree of flexibility, as they can be modified (gaining new details, texts, clippings), saved, and used in different ways, promoting a more personalized and creative expression in chats. That sticker will always be available to the user in their collection unless they choose to delete it.

Finally, while other image media formats have a typical rectangle format, stickers

[2] https://tenor.com/pt-BR/

appear in much more diverse forms, with their edges often shaping the outline of the image they represent and without the need to click to enlarge the content. Unlike images, which users can choose not to show on chat directly, stickers are automatically downloaded and displayed to users in WhatsApp chats [81], making them even more susceptible to abuse by malicious actors. These characteristics make stickers a powerful visual media format, easy to use, and more invasive than others, which may explain why they are being used for attacks within WhatsApp, as reported in Chapter 4.

5.1.1 Stickers Usage Patterns

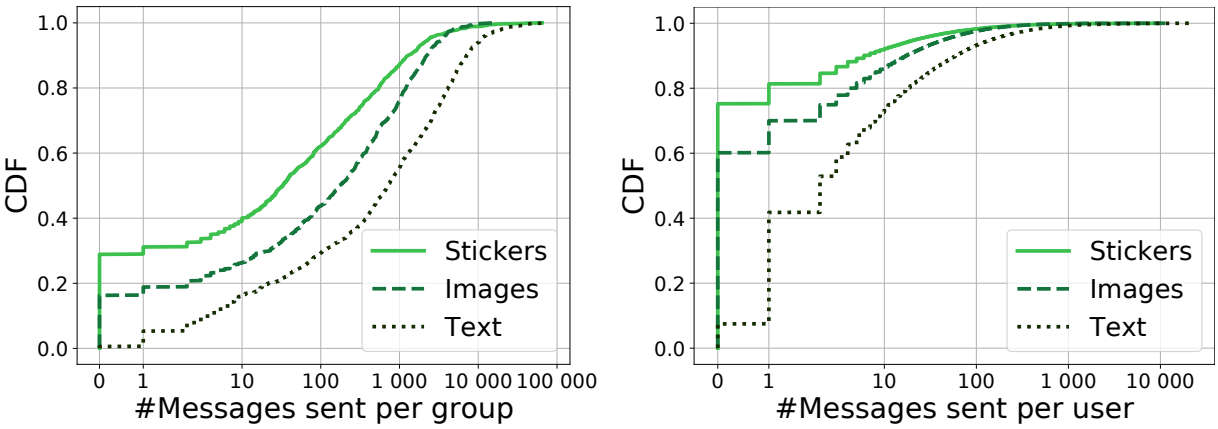
These characteristics, however, refer to the structure and format of the sticker media and not to their usage. We do not know whether the thought process employed by users when sending a sticker is similar to or not from other kinds of media. In this section, we evaluate some aspects of the stickers to investigate how members of the groups use this media.

In Figure 5.1, we compare stickers with images and texts regarding the volume of media sent per group and user. In the cumulative distribution function (CDF) of the messages per group (Fig. 5.2(a)), we observe that sticker messages were less popular in the groups than images. About 30% of the groups did not send any stickers in the chat, images were absent in less than 20% of them, and all groups had text messages. Moreover, 60% of groups sent no more than 100 stickers in total. On the other hand, we note that some groups have more than 10,000 stickers. The less frequent use of stickers may be explained by the political nature of the groups monitored in this study, since some groups expect a more formal communication for the debate while stickers are often associated with a more casual or funny context. In contrast, not all political groups are made only of serious discussions, which can reflect a higher number of stickers.

In the distribution of messages per user (Fig. 5.2(b)), we have similar curves, with stickers being the least common media. Here, even more, users do not send any stickers (75%), but we also have users who individually sent around 10,000 stickers. Hence, we can observe that, as expected, in a messaging app, text format is much more commonly used than other types of media. However, it is interesting to note that users prefer to send images than stickers on the groups, even though both media are widely spread on WhatsApp, especially when considering that stickers are generally easier to send, as they do not need to select a file from the gallery and are more incorporated into the WhatsApp interface.

As we processed and merged stickers, we measured the number of shares each

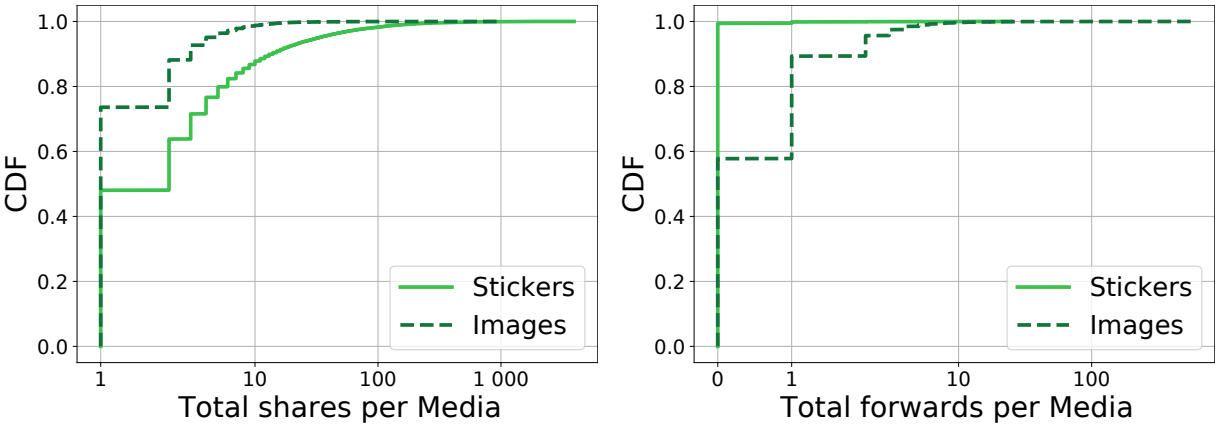
Figure 5.1: Cumulative Distribution Function (CDF) of stickers sent per group and user, compared with image and text.



(a) Groups (b) Users

Source: The Author.

Figure 5.2: Cumulative Distribution Function (CDF) of total shares and forwarding per sticker and image media.



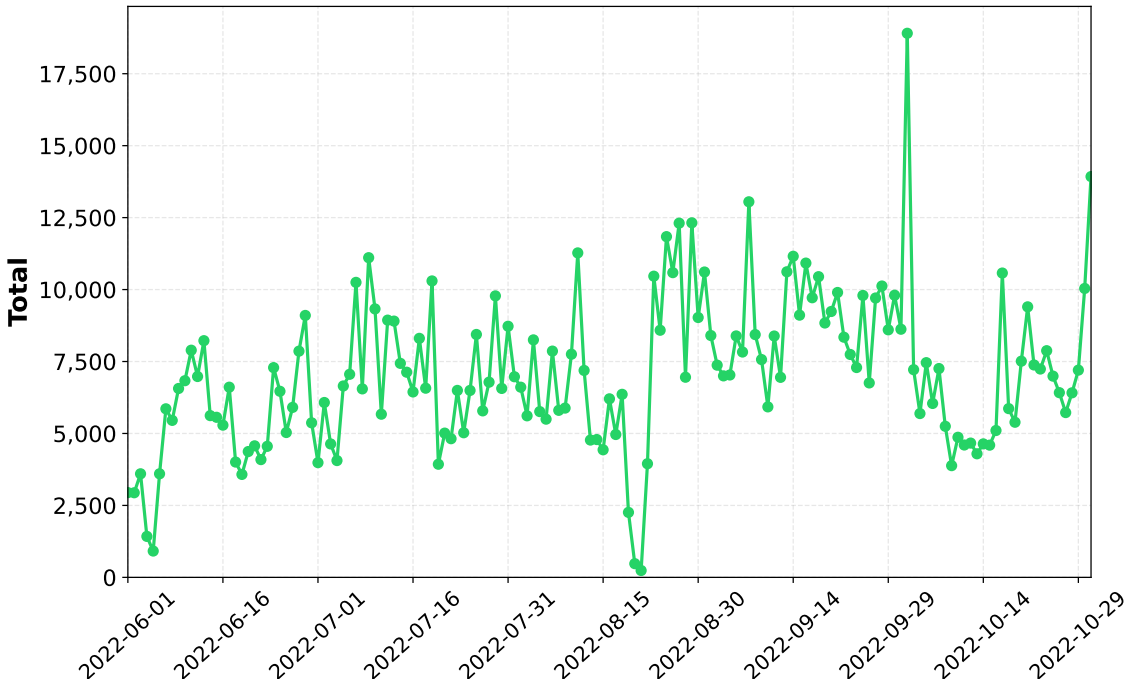
(a) Shares (b) Forwards

Source: The Author.

sticker has within our dataset. In Figure 5.3(a), we evaluate the distribution of total shares per media by comparing images and stickers. Even though we observed more unique image messages in the data, when looking at the total shares per unique media, we note that stickers, in general, have more individual shares than images. About half of the stickers appeared more than once, many of which were shared more than a thousand times. In contrast, about 75% of images have only a single appearance.

Another attribute we obtained for each message is whether it was forwarded or not. Forwarding is an essential tool in the WhatsApp ecosystem, in which users quickly share content with their contacts [102]. A recent study shows that about a quarter of all messages in WhatsApp groups are forwarded [103]. By analyzing forwarded content on our dataset, we found that images are much more frequently forwarded than stickers. Almost none of the collected stickers were found in the forwarded messages, meaning that

Figure 5.3: Stickers sent per day in the WhatsApp dataset.



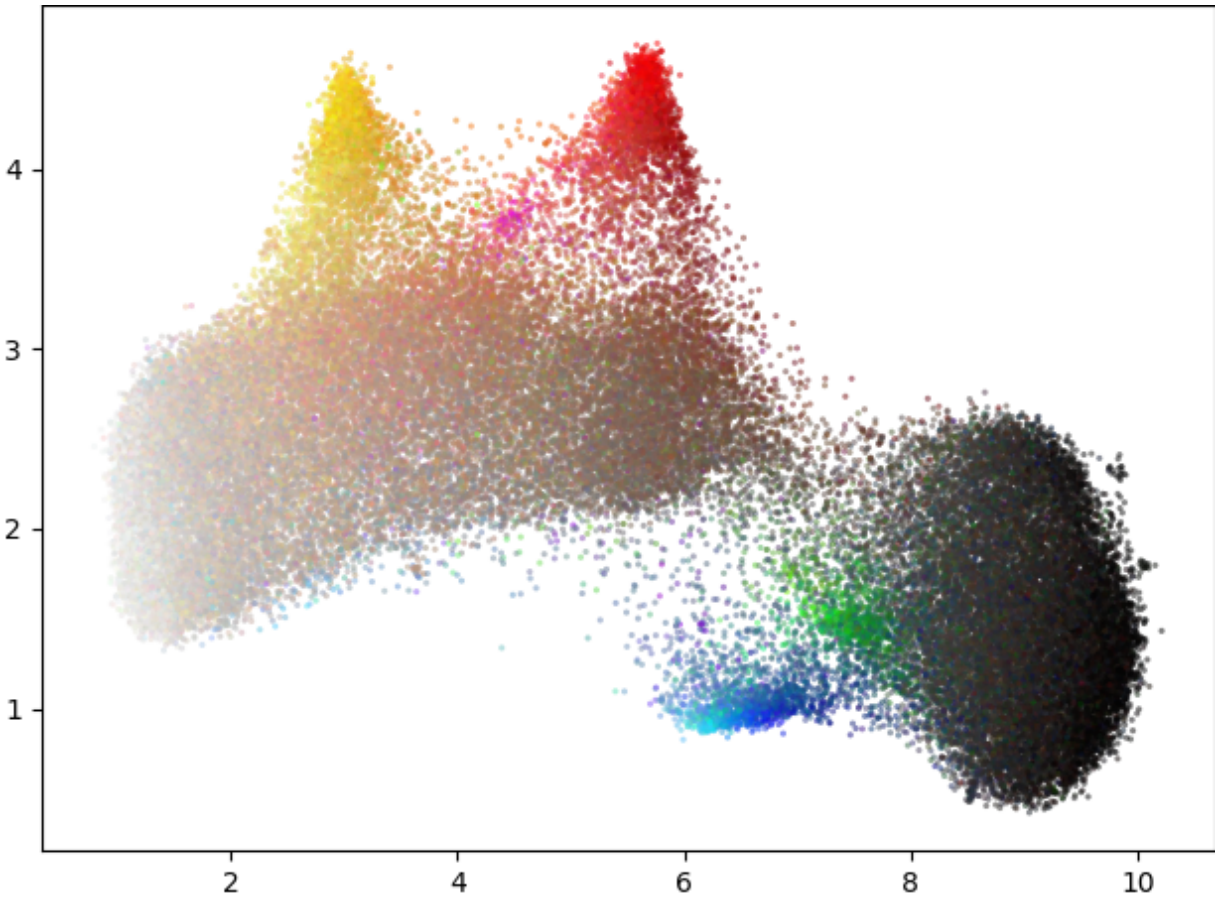
Source: The Author.

nearly all of them were sent directly from the users associated with a group. On the other hand, for images, we observe that 40% have been shared as a forward at least once. This reveals a key difference between these multimedia formats.

Stickers differ from images not only structurally but also in their usage patterns. While there are fewer stickers in total compared to other media types in WhatsApp groups, individual stickers are shared more frequently and directly (not using forwarding) by users in chats. This behavior suggests that stickers may present a collectible quality [170], in which users prefer to curate a collection of stickers that can be readily used in specific situations or as a form of self-expression [154]. To understand this, it is crucial to examine how stickers work on WhatsApp. Without the forwarding mechanism, users need to save them to their collection by marking them as favorite stickers. This process, described in detail by [81], involves selecting a received sticker and choosing the "mark as favorite" option, which adds it to the user's favorites menu. WhatsApp interface reinforces this collection-based experience by providing a dedicated section for users to access their favorite stickers, enabling quick use in chats.

Last, we observed in Figure 5.3 the large volume of stickers posted per day within the public groups monitored from our dataset. Users send over 5,000 stickers per day on average within the groups analyzed, but three significant peaks stand out. The largest occurred on October 6, coinciding with the Brazilian presidential elections, when sticker usage surged to over 17,500, while the second peak was on October 30, during the second round of the elections, with over 13,000 stickers. The third was around September 7,

Figure 5.4: UMAP visualization of all stickers from the WhatsApp dataset.



Source: The Author.

Brazil's Independence Day, an important date in the country's political calendar. It is also interesting to note the gradual increase in the number of stickers as the election period approaches and the sharp drop right after voting day. These spikes underscore the political nature of the dataset and the strong connection between sticker activity and major political events.

5.2 Sticker Content Characterization

In this section, we evaluate sticker content, grouping them by visual similarity, and creating a graph of stickers according to the groups in which they were posted to evaluate.

A visual representation of all stickers collected is shown in Figure 5.4. We created a representation using both the pHash binary vector and each sticker's dominant color. Then, we employed dimensionality reduction via uniform manifold approximation and projection (UMAP) to plot the sticker representations in a 2D space [98]. Each point is

Figure 5.5: Cluster of grouped stickers representing emojis.

(a) (b) (c) (d)

Source: The Author.

associated with the representation of a single sticker, and its color is the dominant color of the actual sticker. It is visible that there is a concentration of red stickers and also green and blue ones, which are related to the context we captured, and related to political campaigns for the 2022 Brazilian elections. The leading left-wing party in Brazil employs red colors associated with the party's branding, while the leading nationalist right-wing party mostly uses green, blue, and yellow colors derived from the Brazilian flag. The stickers in our dataset reflect the partisan nature of the branding of the competing parties. A thorough manual investigation of the yellow chunk of stickers revealed the presence of many "emoji" stickers, which are faces drawn in emoji style representing existing emoji characters. The gray area in the plot encompasses many different sticker styles, and we do not point to a single characteristic of them.

5.2.1 Exploring Visual Similarity

To explore the similarity of these stickers, we leverage image features to cluster and investigate the patterns and characteristics of the stickers. A similar methodology has been employed to study memes in Web communities [167], annotating and mapping them to clusters. Since the perceptual hash (pHash) provides a comparable fingerprint of stickers, we can use it to cluster the stickers through the density-based spatial clustering of applications with noise (DBSCAN) algorithm [35], merging them into visually similar groups of content shared on WhatsApp. Using DBSCAN allows us to investigate the popularity and diversity of stickers and gain insights into their content.

The pHash-grouped sticker clusters reveal sets with variations of images, emojis, or memes, evidencing the dynamics of sticker usage. Some clusters highlight a key characteristic of stickers: their role as both meme-like content and emoji substitutes, as seen in Figures 5.5 and 5.6. Emojis, which are pictorial representations of emotions or objects,

Figure 5.6: Example of the cluster with meme sticker template with small variations.

(a) (b) (c) (d)

Source: The Author.

often take the form of simple, round yellow faces [170]. In the case of stickers, these are frequently more elaborated versions, with some evolving into animated versions. Like emojis, these stickers are commonly used as emotional reactions to messages, providing users with a more visually nuanced way to express themselves in conversations [155].

The meme-associated clusters, on the other hand, consist of a set of analogous images that follow a recognizable template, differentiated only by minor text changes or visual details to express humor or satire. This aligns with the typical behavior of memes, where a base well-known image is edited and remixed across the Internet, maintaining its core template while being adapted to various contexts [55]. This practice of remixing, transforming, and altering base images is central to the meme ecosystem on the Web, where visual elements are modified to reflect cultural or situational humor [86, 87]. Stickers, in this context, also extend this behavior, serving not only as images but as part of a broader visual conversation that blends both expressive and cultural elements.

It is interesting to note that within certain clusters, very visually similar stickers may also portray opposing ideas. In the context of the public political groups encompassed in this work, many clusters are directly related to the poll-leading politicians in the Brazilian 2022 elections and contain advertising/provocative material associated with their campaigns. Here, we can find examples of corresponding stickers that carry drastically opposing partisan content, as shown in Figure 5.7. This cluster grouped two visually analogous stickers, but they are shared in different contexts. The stickers portray an edited image of opposing candidates of the left and right-wing parties in a criminal photo, suggesting that stickers are co-opted by adversary groups and used with opposing semantics but adopted with the same visual aesthetics. These results alert us to the implications of grouping stickers (or any image content) exclusively based on their visual appearance. Small details often imply drastic changes in the semantic value of the sticker.

Figure 5.7: Examples of visually similar stickers used by opposing political leanings.

(a) (b)

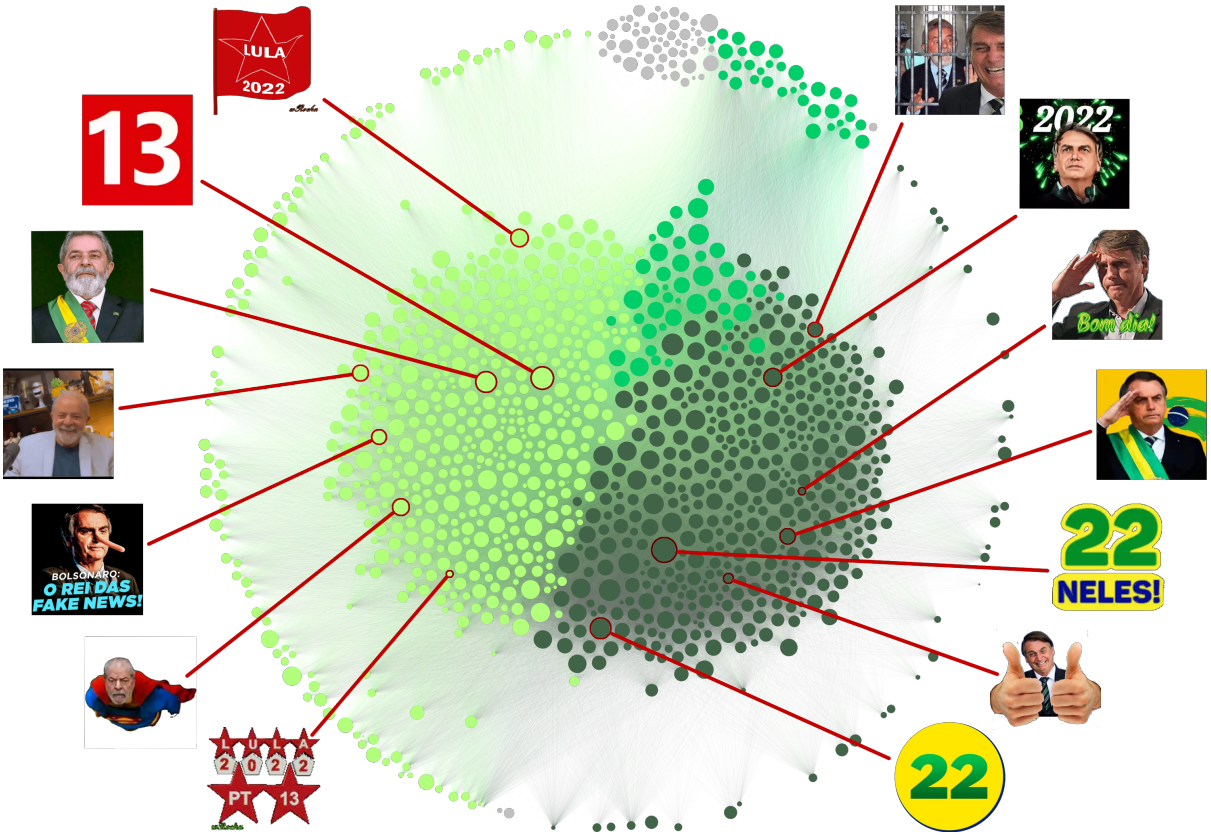
Source: The Author.

5.2.2 Political Alignment of Stickers

As exemplified in the previous section, visually similar stickers can convey drastically different ideas and be used in entirely different contexts. Since groups analyzed in our dataset expose differences in political alignment, activism level, or even in goals or topics discussed, besides the visual appearance of stickers, we also evaluate their similarity based on the context in which they are used. To evaluate the content of stickers from this perspective, and check whether stickers sent in similar groups express similar ideas, we build a graph of stickers shared within the same chats to create a network of stickers based on the groups in which they were shared. Figure 5.8 offers a graph visualization of this network assembled with the most popular stickers in our WhatsApp data. In this graph, each node is a sticker, and they are linked with an edge if both appear in the same group. We also applied a community detection algorithm based on modularity optimization [12] to identify sticker communities within the network. For visualization clarity, we used only stickers shared more than 100 times in the observed groups and edges with at least 10 groups in common.

Even with a very dense network (density measured at 0.698), we could identify two main larger communities and a smaller one. These two main communities reflect the dataset's polarization, featuring very similar stickers from both political leanings. There were two major candidates for the Brazilian 2022 presidential elections (Bolsonaro, a right-wing candidate, and Lula, a left-wing candidate). Most of the stickers found in the monitored public groups represent both sides of this dispute. While some stickers are symbols, numbers, draws, and colors associated with each political party, others portray both candidates as being in a position of leadership as president or in a position of defeat.

Figure 5.8: Stickers network.



Source: The Author.

Numerous stickers aim to criticize or anger members of the opposing political leaning. Together, these sticker archetypes form a unique and connected political community. In contrast, the smaller third community features more diverse content, including sexual content and generic stickers. To analyze the partisan leaning of stickers, we label each monitored group as left-wing, right-wing, or undefined, based on the political stance indicated in the group's name and, when needed, its description. The group annotation shows that most groups support the right-wing candidate Bolsonaro (666 groups), while only 126 groups support a left-wing agenda. The remaining 617 groups were tagged as undefined, as their political alignment could not be determined (e.g., general debate groups or support for other political positions).

Next, we analyze each sticker to determine its political leaning based on the groups that shared it. Then, using the labels of these groups, we calculate the proportion linked to each political side and assign the sticker a label based on the side with the highest proportion. This label indicates whether the sticker is primarily used in right-wing or left-wing contexts. Note that some stickers may be used equally by both sides of the political spectrum or predominantly by groups without a clear political alignment. To prevent mislabeling in such cases, we introduced the third category, "undefined", for stickers without at least 60% majority on either side. This approach allows us to better

	Right	Left	Undefined
Total Groups	666	126	617
Total Stickers	18,158	6,704	32,169

Table 5.1: Political alignment of groups and sticker context.

identify stickers that are closely linked to a political alignment. Through this annotation, we found that 18,000 stickers are strong biased towards usage in right-wing groups and 6,000 towards the left, as shown in Table 5.1. Note that the difference could be explained by the fact that, in Brazil, the right-wing movement is seen to be actively mobilized on WhatsApp [16].

5.3 Political Attacks with Stickers

Given the potential for stickers to convey strong partisan sentiments, our analysis delves into the scenarios in which political stickers serve as tools for provocation or even direct attacks against opposing political groups using our methodology. In the Chapter 4, we analyzed political activism on WhatsApp in Brazil, revealing common types of attacks performed by partisan public groups towards opponent campaigns on WhatsApp. The public nature of these political groups easily allows individuals from adversarial political positions to join, establishing a channel for launching targeted assaults against individuals aligned with opposing ideologies. The ease of one-click sticker sharing may explain why these initiatives are particularly conducive to flooding group chats with divisive content. By identifying the political leaning of stickers, we can examine whether they are being misused in public groups by political campaigns to harass opposing users. To investigate this, we analyze whether stickers predominantly associated with a specific political alignment appear in groups aligned with the opposing spectrum.

Our analysis shows that most politically annotated stickers were used exclusively within their respective ideological groups. However, we also found cases where political stickers transcended party boundaries, appearing in groups with opposing leanings, suggesting politically motivated attacks. Based on these criteria, we identified 869 partisan stickers, with a notable majority (794) exhibiting right-wing bias and being shared in leftist groups. In contrast, only 75 left-wing biased stickers appeared in right-wing groups. A manual evaluation revealed that many of these stickers are either political propaganda supporting the opposing candidate or "humorous" memes intended to provoke group members. We also observed stickers portraying both candidates in derogatory and

humiliating ways, including edited images with sexual connotations, highlighting the offensive nature of some attacks. Notably, creating WhatsApp stickers using a person's face without permission, particularly when intended to insult, harass, or damage their reputation, is already considered illegal in countries like Indonesia [84]. In Brazil, this issue is regulated by broader defamation laws, and there is no specific regulation for WhatsApp. Nonetheless, the Superior Electoral Court has criminalized the massive spread of political content through message applications.[3] In Germany, police reported that child sexual abuse stickers were shared in a climate activism group, putting all members at risk of having illegal content on their devices [81].

Figure 5.9 shows examples of these "political attack" stickers, which promote a partisan candidate in rival environments. The presence of such stickers in politically opposing groups suggests a deliberate act of confrontation by users, trying to cause reactions and distress among individuals holding contrasting political views. For instance, Figures 5.10(a), 5.10(b), 5.10(c), and 5.10(d) depict stickers featuring the right-wing presidential candidate Bolsonaro, perceived as provocative within leftist groups. One sticker even includes the text "Leftists will infarct with hate", tailored to instigate distress. Similarly, leftist stickers depicted in Figures 5.10(g) and 5.10(h) endorse Lula and are also incongruous with right-wing ideologies. Of particular concern are images such as those in Figure 5.10(e), depicting Bolsonaro wielding a firearm, and Figure 5.10(f), featuring heavy boots crushing a red star with the words *"This is our fight. Communism, not here!"*. These images evoke menacing connotations, thereby exacerbating tensions within a polarized online political discourse.

Our findings reveal that clustering stickers based on their image pHashes, which denotes visual similarity, enables the identification of image sets exhibiting even minor variations. However, this may not capture the subtle political context carried with them and may also inadvertently put together stickers with opposite political messages. On the other hand, examining stickers through the perspective of the groups in which they are shared, and thus their political alignment, exhibits a strong association with the contextual environment rather than solely relying on visual resemblances of the images. This highlights the importance of considering the broader sociopolitical context in understanding sticker usage patterns. Furthermore, we observed that stickers often demonstrate remarkably partisan alignments and are frequently utilized as tools for provocation and abuse within political public groups on WhatsApp. This highlights the pivotal role of stickers in expressing and perpetuating political ideologies and tensions within this network.

Moreover, the asymmetric distribution of stickers aligned with right-wing ideologies may reinforce echo chambers and polarization within WhatsApp groups. Users who predominantly encounter content aligned with their own political beliefs may become

[3] https://folha.com/zdu068gh

Figure 5.9: Example of stickers used to "provoke" or "attack" opposing political groups.

(a) (b) (c) (d)

(e) (f) (g) (h)

Source: The Author.

further entrenched in their viewpoints, hindering dialogue and deepening division in online communities. Sharing biased stickers also raises important questions about political engagement and manipulation on social media platforms. The deliberate dissemination of partisan content, particularly in the form of stickers designed to provoke or attack opposing political groups, shows a strategic use of digital media for political propaganda.

5.4 Abusive and Hate Stickers on WhatsApp

Although politically motivated attacks are one form of sticker abuse, others involve offensive and inappropriate content on WhatsApp. In this section, we delve deeper into these pathways, shedding light on the various forms of abusive behavior facilitated by stickers on the platform, including the spread of hate speech. We use the NSFW Yahoo model [91] to measure information on how "explicit" each sticker is. We also use SafeSearch detection with Google Vision API[4] to get complementary data about potentially harmful content being posted on WhatsApp through the stickers.

On WhatsApp, users can freely create customized stickers based on any image they want. However, this model is not unanimous among all messaging apps. Most

[4] https://cloud.google.com/vision/docs/detecting-safe-search

of them have only predetermined and limited sets of stickers that users can use during conversations. Facebook and its messaging system "Messenger" do not allow any users to add their own creations as stickers [44]. The mobile messaging app LINE even sells millions of curated sticker packs per month, and stickers have long been one of LINE's key revenue drivers [131]. Despite Meta restricting stickers on its other platforms and selling stickers can be a highly profitable economic model, for WhatsApp, particularly, Meta chose to let users freely create and share their own stickers. However, this permissive model has some challenges, as it opens the way for the dissemination of offensive and inappropriate content. Not surprisingly, stickers are used to distribute not only memes, but also illegal content such as child sexual abuse material [43] and Nazi propaganda [136]. Even worse, WhatsApp automatically saves every sticker received in a chat, hence users may have incriminating stickers on their devices without knowing or wanting to [81].

WhatsApp's terms of service, however, state that stickers created must be legal, authorized, and acceptable images. Furthermore, users are not allowed to use WhatsApp services in ways "*that are obscene, defamatory, threatening, intimidating, harassing, hateful, racially or ethnically offensive, or instigate or encourage conduct that would be illegal, such as promoting violent crimes*". Although WhatsApp's TOS does not allow for offensive stickers, a quick manual inspection of the top 5,000 most popular stickers shared within our dataset revealed clear examples of stickers that do not comply with these rules. There were instances of stickers depicting extreme violence, hate symbols, degrading pornography, and highly repulsive images.

Figure 5.10 presents hate stickers found in our dataset. Sticker 5.11(a) is an example of a homophobic image against LGBTQIA+ people; Sticker 5.11(b) depicts a swastika, symbolizing Nazism; Sticker 5.11(c) shows a black man holding a knife alongside the phrase "around blacks never relax", which is a recurring racist remark on the Web. Similarly, Sticker 5.11(d) portrays a derogatory caricature of a Jewish man, reflecting antisemitic stereotypes, which is also a widely known hate symbol.[5] These examples demonstrate that stickers, freely sent and shared between users on WhatsApp, can be used as media for targeted harassment and attacks against marginalized communities.

In our data, we also found a considerable presence of stickers with sexually explicit content. To evaluate this, we apply the convolutional neural network model for Not Safe for Work (NSFW) proposed by Yahoo Inc. [91] to identify adult-themed stickers. Images with a score greater than 80% are labeled as explicit. Figure 5.12(a) presents the volume of NSFW stickers. Compared to images, stickers are five times more likely to depict explicit content (0.5% of images and 2.3% of stickers are NSFW). In total, we discovered 33,335 messages containing NSFW stickers, accounting for 5.5% of all sticker messages. As shown in Figure 5.12(b), around 10% of users in the monitored groups shared NSFW sticker, and some individuals posted hundreds of NSFW stickers. There is even a peculiar

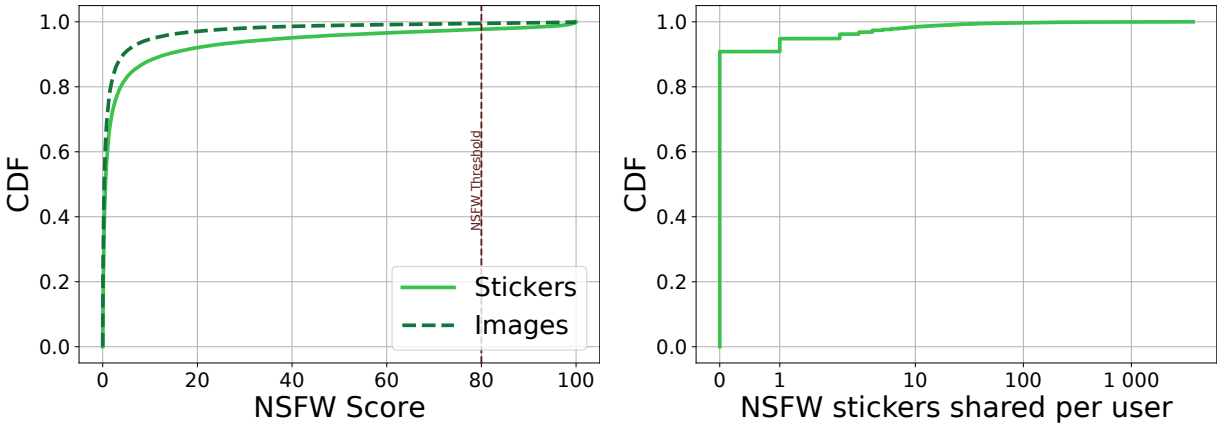
[5] https://www.adl.org/resources/hate-symbol/happy-merchant

Figure 5.10: Example of offensive stickers sent by users that violate WhatsApp's terms of use.

(a) (b) (c) (d)

Source: The Author.

Figure 5.11: Distribution of NSFW stickers sent.



(a) NSFW Score (b) NSFW per Users

Source: The Author.

user who shared more than 3,000 NSFW stickers. These findings suggest abusive behavior, particularly given that these public political groups are typically not intended for adult content, with some even explicitly forbidding it in their descriptions.

Furthermore, we evaluate other categories of NSFW content through the Safe Search detection from Google's Vision API.[6] This tool detects content within an image across five categories (adult, spoof, medical, violence, and racy) and returns the likelihood for each one of them. According to the API, Adult content encompasses elements such as nudity, pornographic images, cartoons, and activities of a sexual nature. The medical category accounts for the likelihood that the content is associated with medical imagery (e.g. wounds, surgeries, blood). Violence offers an estimation of the probability that the image portrays a violent action. Spoof is the likelihood that a modification was made to the image's canonical version to make it appear funny or offensive. Content deemed racy may include (but is not limited to) skimpy clothing and strategically covered nudity, lewd, provocative poses, and extreme close-ups of sensitive body areas. In Table 5.2, we aggre-

[6] https://cloud.google.com/vision

	Unique Stickers	Number of Shares
Yahoo_NSFW	1,322 (2.3%)	33,335 (5.05%)
Adult	1,423 (2.5%)	30,828 (4.6%)
Violence	416 (0.7%)	4,583 (0.7%)
Racy	4,503 (7.9%)	55,476 (8.4%)
Medical	735 (1.3%)	14,300 (2.1%)
Spoof	17,834 (31.2%)	150,351 (22.7%)

Table 5.2: "Not Safe" stickers categories on WhatsApp.

gated the results of Safe Search detection on our data collection of stickers, considering the likely and very likely labels for each category.

These findings underscore the urgent need for moderation and enforcement of content guidelines within WhatsApp's sticker ecosystem. While stickers enhance user engagement and communication [154], their misuse brings significant risks, including promoting hate speech, reinforcing harmful stereotypes, and creating unsafe environments within online communities. Our analysis suggests adopting strategies such as using automated tools to detect offensive, NSFW, or adult content and restrict potentially harmful stickers. Despite WhatsApp's encrypted ecosystem, studies have proposed system architectures for moderate content on the platform [77] without violating the privacy of its users [122]. Since these stickers are primarily derived from existing images, the platform could offer alternatives to restrict the creation of content that violates community standards.

5.5 Summary

In this chapter, we explored the multifaceted role of stickers in WhatsApp's political ecosystem, focusing both on their expressive potential and on their abuse during the 2022 Brazilian presidential election. Analyzing 650,000 sticker messages across 1,600 public groups, we characterized their usage patterns, visual features, political alignment, and potential for abuse. Our findings reveal some characteristics that suggest stickers are indeed a distinctive media format while also sharing similarities with others: stickers are less frequently sent compared to other media types, such as images within groups. However, a single sticker usually presents higher total shares, which means they exhibit a higher rate of recurrence within conversations, similar to an emoji. Moreover, stickers are rarely forwarded, suggesting that users often save and keep their collection of preferred stickers for future use instead of using the system's forwarding tools. This behavior highlights the collectible nature of stickers, in which individuals accumulate a personalized

repertoire of stickers to deploy on various occasions, like an emoji or a meme.

However, we also uncover instances where visually similar stickers convey entirely different meanings, emphasizing the importance of considering other aspects when grouping images by perceptual hashes. Particularly in political groups, where misinformation is a concerning issue, it is important to avoid merging distinct ideas within the same category, especially when minor alterations in images could propagate misleading narratives. Furthermore, we analyzed the political alignment of the stickers by building a network of co-occurrence in public groups. This network analysis resulted in two prominent communities of stickers, mirroring the polarized partisan landscape of our political WhatsApp dataset, and finding a strong association between stickers and political affiliations. Notably, we observe instances where political stickers are shared across ideologically opposing groups, often serving as tools for attacks and abuse on WhatsApp.

Our results reveal the urgent need for better moderation mechanisms to curb the dissemination of offensive stickers and safeguard users from encountering inappropriate content on WhatsApp. We identified a significant prevalence of potentially offensive sticker messages, including homophobia, hate symbols, racist stereotypes, and derogatory depictions of marginalized groups. We also observed that stickers are five times more likely to depict explicit (NSFW) content compared to images. Explicit content accounts for 2.3% of unique stickers and 5% of all sticker messages. Additionally, we identified a substantial volume of violent (0.7%), medical (2.1%), and racy (8.4%) stickers, which may be sensitive for certain audiences, especially in public groups. Given the visual immediacy of stickers, the unrestricted access to public groups, and the ease with which users can create and disseminate stickers from virtually any image, there are significant risks of offensive content spreading unchecked due to a lack of moderation. Despite WhatsApp's private and encrypted nature, our research provides valuable insights into how stickers are used within public political groups, bringing more understanding and transparency to the platform, particularly about the substantial abuse of stickers with offensive content disseminated on WhatsApp, which requires actions to bring more safety to users.

Chapter 6

From Fake News to Real Protests: Coordination in Public Groups

Upcoming elections in different countries come together with serious concerns about election integrity. Those concerns are mainly associated with the uncontrolled dissemination of misinformation in social media [60], the increasing polarization [27] and radicalization [129], the indiscriminate use of targeted advertising [137] and social bots [46], the increasingly personalized feed algorithms from social media platforms [129]. Those concerns are more worrisome as different AI models become readily accessible, simplifying the creation of misinformation [7].

Since 2018, misinformation campaigns in Brazil took place in a new digital space, yet poorly understood: messaging platforms such as WhatsApp [10]. After the 2018 elections in Brazil, WhatsApp acknowledged the existence of coordinated campaigns that spread massive amounts of messages during the 2018 presidential elections in Brazil [99]. To mitigate the problem, WhatsApp took steps to reduce its virality features by limiting how much content can be forwarded [102]. On the other hand, the Superior Electoral Court, responsible for Brazil's elections, has criminalized the massive spread of political content through message applications[1]

However, although those countermeasures are very welcome, they are still limited. First, bypassing the forwarding limit was quite simple and ineffective [103]. Second, WhatsApp introduced new features that increased virality and facilitated massive spreading. For example, WhatsApp introduced communities, which allow one to manage and post in multiple groups simultaneously.[2] Finally, although a coordinated campaign for sharing political content in WhatsApp can be considered an electoral crime in Brazil, auditing any activity in WhatsApp is very difficult, given the closed nature of the application [120].

In this chapter, we aim to investigate the existence of coordinated campaigns. Specifically, we will examine whether there is evidence of coordinated accounts actively spreading messages on the platform. Furthermore, we will explore the content and goals of the coordinated messages and analyze how they relate to recent Brazilian political events.

[1] https://folha.com/zdu068gh
[2] https://faq.whatsapp.com/495856382464992

Despite there are works studying coordination on WhatsApp [114, 115], they bring a definition of coordination at the group level, using network structure and backbone extraction to identify coordination by the similarity of the content posted. In contrast, our paper presents a different and more restrictive definition, rapid coordination, adapted from [118]. This approach considers the content similarity and the synchronous posting behavior to identify coordinated accounts. This brings a completely different perspective on coordination in WhatsApp, which has not been explored before. Here, we focus on identifying consistent and synchronous coordination efforts on WhatsApp that leverage the instantaneous nature of the platform to boost and amplify messages. Furthermore, to the best of our knowledge, this is the first large-scale study to explore rapid coordination on WhatsApp using a large dataset of more than 13M messages collected over seven months, covering recent important events in Brazil. We gathered an extensive messaging collection of Brazilian political public groups, considering seven months of data from July 2022 until January 2023. This period captures significant Brazilian events such as the 2022 presidential election, attacks on the electoral process, and riots that ended with an attack on Brazil's federal government buildings on 8 January[3]. In total, this study analyzes 13,452,039 million messages shared in 1,444 groups from 100 thousand users. The huge data volume allows us to observe a different perspective by incorporating temporal similarity to identify synchronous coordination actions. By including a similarity of time posting, we can identify more consistent cooperation between accounts [58]. Additionally, we aim to investigate different coordination formats, including text, images, and videos.

This chapter is organized as follows: First, we define rapid coordination and present the modeling approach used to identify such coordination (Section 6.1). Next, we apply this approach to detect coordination activity (Section 6.2) and analyze coordination messages (Section 6.3), characterizing the text (Section 6.3.1), URLs (Section 6.3.2), and images (Section 6.3.3) involved. Finally, we explore the content of these coordination messages to better understand the goals behind the coordinated efforts (Section 6.4).

6.1 Rapid Coordination

Coordinated activity is a well-known phenomenon on various social networks, where users employ it for various purposes, such as sharing beliefs, marketing, mobilizing people, shaping public opinion, and spreading misinformation [112]. In WhatsApp, the main messaging app in Brazil, particularly for political purposes, coordinated accounts can take advantage of the platform to amplify engagement in specific activities.

[3] https://en.wikipedia.org/wiki/2023_Brazilian_Congress_attack

These groups of users often replicate similar behaviors over time, such as endorsing certain messages and amplifying specific content.

Due to the widespread use of WhatsApp for content spreading, our goal is to identify networks of accounts that share the same content rapidly, simultaneously, and repeatedly in a coordinated way. Although there are works studying coordination on WhatsApp [114, 115], they bring a definition of coordination to the group level, using network structure and backbone extraction to identify coordination by the similarity of posted content. In contrast, our goal represents a different and more restrictive definition, rapid coordination. This approach considers the content similarity and the synchronous posting behavior to identify coordinated accounts. This brings a completely different perspective on coordination in WhatsApp, which has not been explored before. Here, we focus on identifying consistent and synchronous coordination efforts on WhatsApp that leverage the instantaneous nature of the platform to amplify messages.

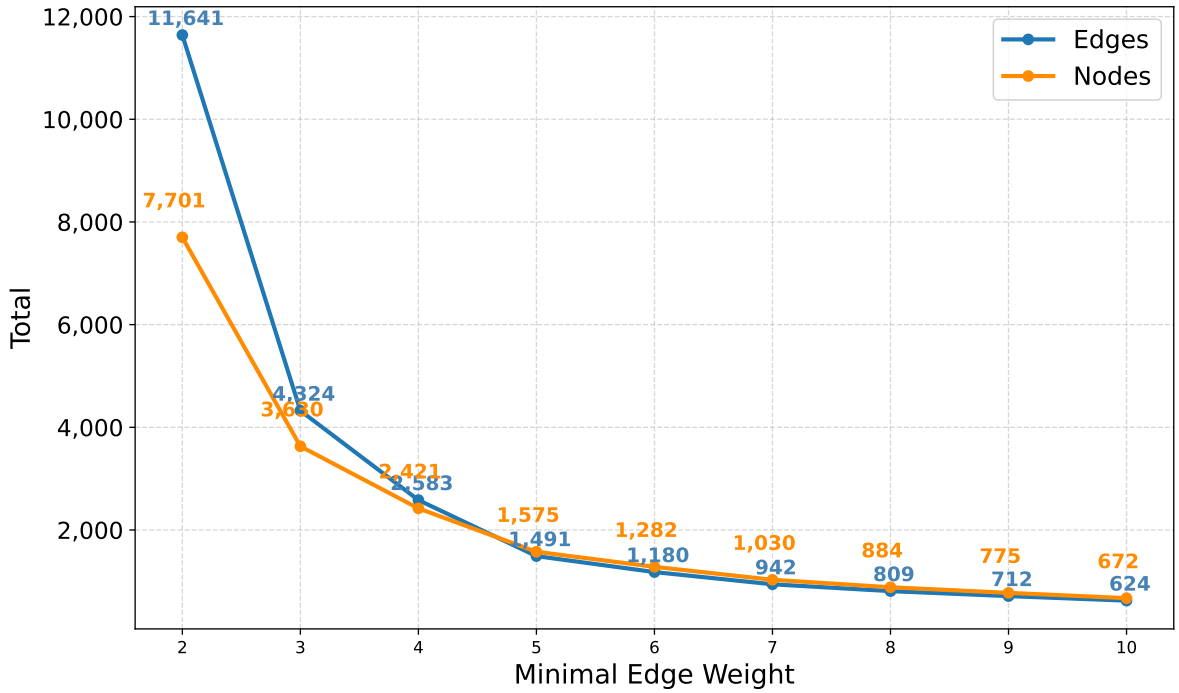
6.1.1 Rapid Spread Network Modeling

To achieve this, we adapted the Rapid Retweet Network proposed by [118], which was originally applied to the Twitter ecosystem, focusing on retweet actions. This approach is particularly effective for identifying groups of accounts that consistently retweet the same source. On WhatsApp, there are no retweets like on Twitter. Instead, users typically receive content and share it in two main ways: directly forwarding the message to another group or copying and pasting the content into a different group.

We define the Rapid Spread Network based on the simultaneous sharing of similar messages within a specific time window. To analyze content similarity within WhatsApp data, we adopted the MD5 hash algorithm to detect identical messages and identify shares of the same content. This method assigns a unique identifier to each unique message, where two messages are considered identical only if they have the same hash. Any modification results in a different hash. With these definitions, we build a weighted network where users are connected if they post the same message in the same time window. In the network $G(V, E)$, a node $v \in V$ corresponds to users who post messages on WhatsApp. The undirected weighted edge $e = (v_i, v_j)$ is included in the users v_i and v_j if they posted the same message in the same time interval. The edge weight of w_{ij} represents the total number of messages they have shared in common during the time window.

On WhatsApp, the private nature of the platform prevents the identification of the source or promoters of a message. Consequently, the network analyzed in this study is composed of users who disseminate identical content simultaneously, highlighting the

Figure 6.1: Coordinated users by alterations in parameters.



Source: The Author.

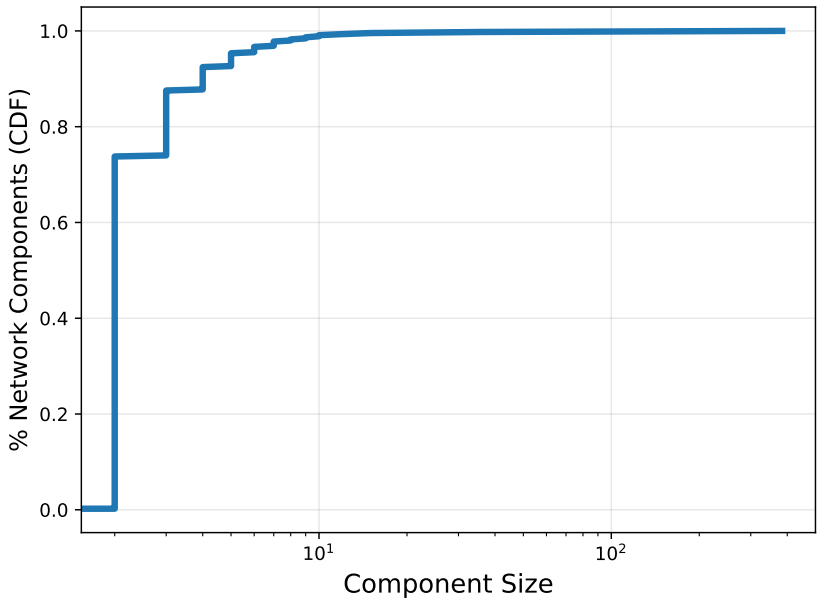
accounts and messages propagated concurrently across the network.

6.2 Identifying Coordinated Activity

Before applying the Rapid Spread Network to model the propagation of coordinated messages, we implemented a series of restrictive adjustments and parameter selections to ensure that we would only capture orchestrated actions by coordinated accounts. First, we selected a 60-second time window[4] to define rapid actions, allowing us to identify potential coordination by users acting simultaneously [78]. To avoid misclassifying common messages frequently shared on WhatsApp (e.g., "good morning", "hello") as coordinated actions, we also filtered out short messages with less than two words. Given the message flow and the large period observed, we tested various threshold values, as shown in Figure 6.1. Using the elbow method, we determined the optimal threshold by identifying the inflection point on the curve. Consequently, we filtered out edges with weights below five to build the final graph. The resulting coordination network comprises 1,575 nodes and 1,491 edges, with an average degree of 1.89.

[4]We performed a sensitivity analysis with varying thresholds to determine whether 60 seconds is a suitable threshold. Please see Appendix B for more details.

Figure 6.2: Component size distribution.

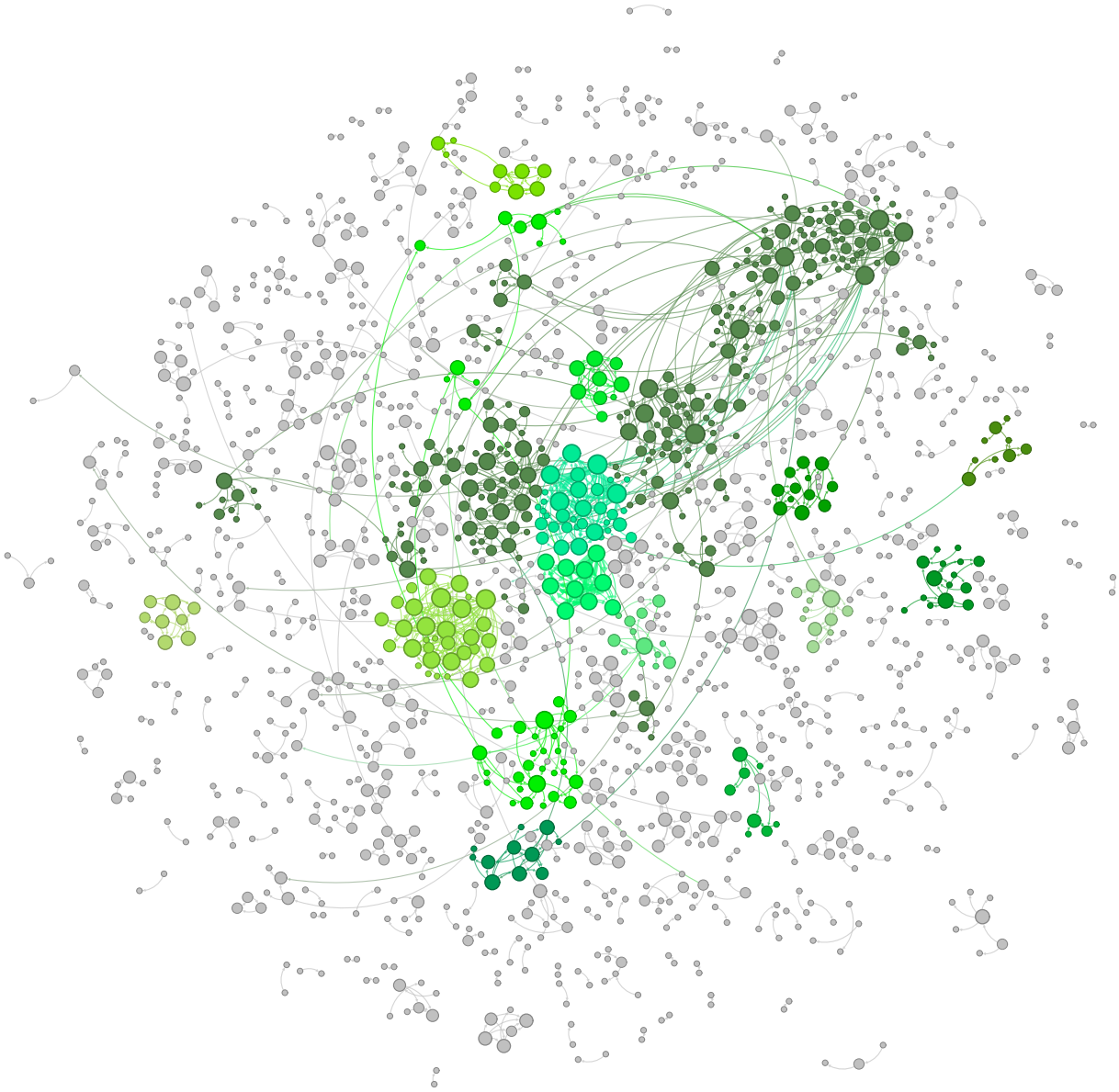


Source: The Author.

Although we are only examining a subset of the WhatsApp ecosystem, we can still observe a clear structure in the propagation of messages. One notable feature of this network is the presence of 450 unique components, which suggests that the network is not densely connected. This fragmentation is significant because it reveals the existence of many isolated pairs or small groups of accounts acting independently to coordinate messages. This high number of unique components suggests a decentralized coordination effort on WhatsApp, where many accounts operate separately to amplify their content rather than a single cohesive group controlling the flow of information. Figure 6.2 presents the cumulative distribution function (CDF) of component sizes, showing that the majority of coordinated components consist of two accounts (73.7%).

Furthermore, we identified a large connected component comprising 332 nodes (21% of the total network). This more connected group demonstrates that while much of the network consists of isolated actors, there is another aspect of coordination, with a substantial cluster of coordinated accounts operating together. This large component suggests a more organized structure within the WhatsApp ecosystem, in which many accounts are coordinated to propagate messages quickly and efficiently. Given WhatsApp's closed and encrypted architecture, it may be challenging for authentic users to differentiate this coordinated activity, making it even more concerning. Using the Louvain community detection algorithm [13], we further identified some communities within the large connected component, highlighting the coordinated nature of these accounts, as shown in Figure 6.3. The presence of multiple communities suggests that coordination on WhatsApp can extend beyond pairs of synchronized accounts, allowing them to reach a wider

Figure 6.3: Rapid Coordination Network.

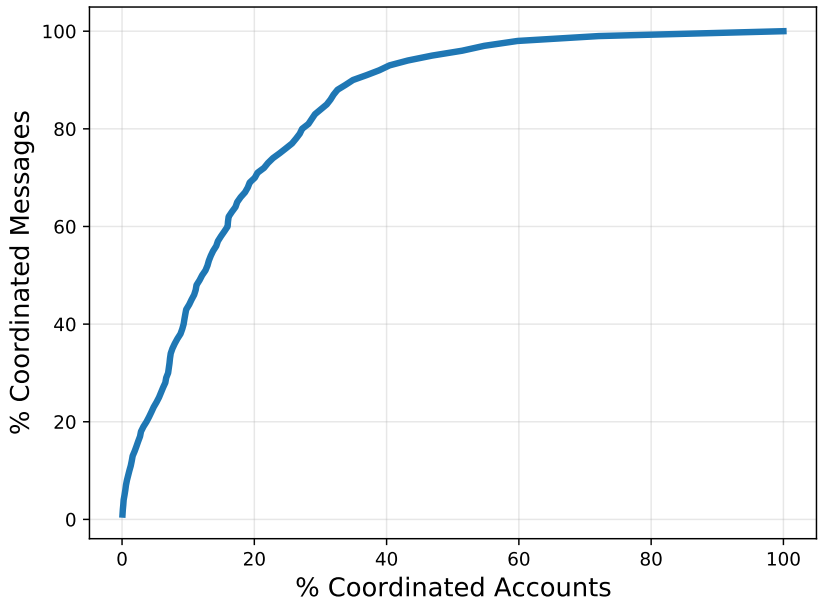


Source: The Author.

audience. These communities could be used to amplify specific narratives or target particular themes within political discussions. We observed that the three largest identified communities contribute significantly to the flow of message propagation on WhatsApp. These communities consist of 301 accounts (19% of all coordinated accounts) and together posted 4,982 messages (34.5% of all coordinated messages). Coordinated accounts in these communities reached 664 groups (45.9% of all groups observed). This scenario helps us understand the impact of coordination and how the WhatsApp environment is conducive to coordinated actions, in which some users can affect a representative part of the group ecosystem.

After we identified the coordinated accounts, we analyzed their messages. Ac-

Figure 6.4: Coordinated messages by coordinated accounts.



Source: The Author.

cording to our definition, messages are considered coordinated only if they are posted simultaneously by two or more coordinated accounts in the same time window. Considering these synchronous coordinated messages, we evaluated the portion of messages posted by the coordinated accounts within our dataset. We identified a concentration of coordinated activity, in which we observed that 80% of the coordinated messages are generated by 27.2% of the accounts, as shown in Figure 6.4. This suggests that, while many users are involved in sending messages, a relatively small subset of coordinated accounts is responsible for most of the messaging campaigns within the political public groups. This indicates a structured and orchestrated use of WhatsApp for coordinated political activity, with some actors playing a key role in shaping the network dynamics. However, this remains largely hidden from public view due to the platform's design.

Takeaways The main takeaways from this section are:

- Our methodology has identified more than 1,575 coordinated accounts actively disseminating political messages.
- There are decentralized coordination efforts in which accounts operate separately to amplify their messages, totaling 450 components.
- A small part of coordinated accounts (27.2%) are responsible for most of the flow of coordinated messages (80%).

6.3 Analyzing Coordinated Messages

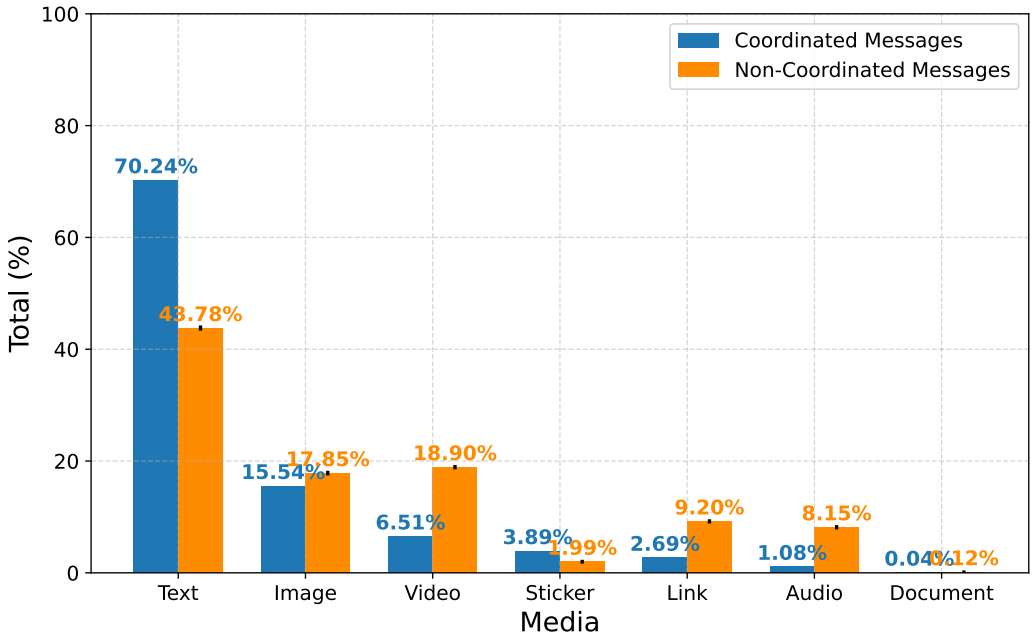
To delve even further into coordinated accounts, we examined their activities, focusing on understanding the messages they shared and their underlying motivations. Since we identified the coordinated accounts, we identified their posted coordinated messages, totaling 14,414 messages. It is important to remember that based on our coordination definition, a coordinated message is posted simultaneously by two or more accounts.

To provide a more contextualized analysis, we created 35 random samples of non-coordinated messages, each matched in size to the coordinated messages. This allowed us to perform a comparative analysis to identify distinctive patterns between coordinated and non-coordinated messages, while ensuring the robustness of our findings with a 95% confidence interval. Similarly to rapid network construction, this sample only includes text messages with more than two words. Figure 6.5 compares the types of media found in these two groups of messages. Text messages are the most common form of communication in coordinated messages, comprising 70.24% (10,124 messages) of the total 14,414 coordinated messages, compared to 43.78% ($\pm$0.1702) in non-coordinated. This significant difference highlights the efficiency of text messages, which are easily shared and forwarded without access to external media files or galleries. Images appeared in 15.54% of coordinated messages, aligning with 17.85% ($\pm$0.1145) in non-coordinated messages. Notably, videos, links, audio, and documents are rarely used in coordinated messages.

Another notable observation is that stickers are more common in coordinated content. While stickers are generally used as a spontaneous form of communication, their use among coordinated users suggests that they can serve as a resource to a flooding attack in a group [74]. A flooding attack occurs when one or more users overwhelm a group with a high volume of duplicate messages in a short period, intending to disrupt the flow of conversation or even crash the group, and usually use stickers [74]. When we applied this definition to observe one of the most popular right-wing groups in our dataset, we identified a flooding attack in which three coordinated users sent over 1,200 duplicate sticker messages within seven minutes. These stickers were primarily composed of provocative content and political attacks. This suggests that, although stickers are less commonly used than text messages, coordinated users can significantly amplify the impact of a sticker flooding attack, making it more effective in achieving its disruptive goals.

When analyzing coordination efforts, it becomes clear that the main goal is to spread messages among as many groups as possible. Text messages are particularly well-suited for this purpose, as they are easy to share and read. Unlike other media formats that require downloads or additional actions, text messages allow for direct and easy communication. This simplicity makes them an ideal choice for disseminating coordinated

Figure 6.5: Messages media type differences from coordinated and non-coordinated messages.



Source: The Author.

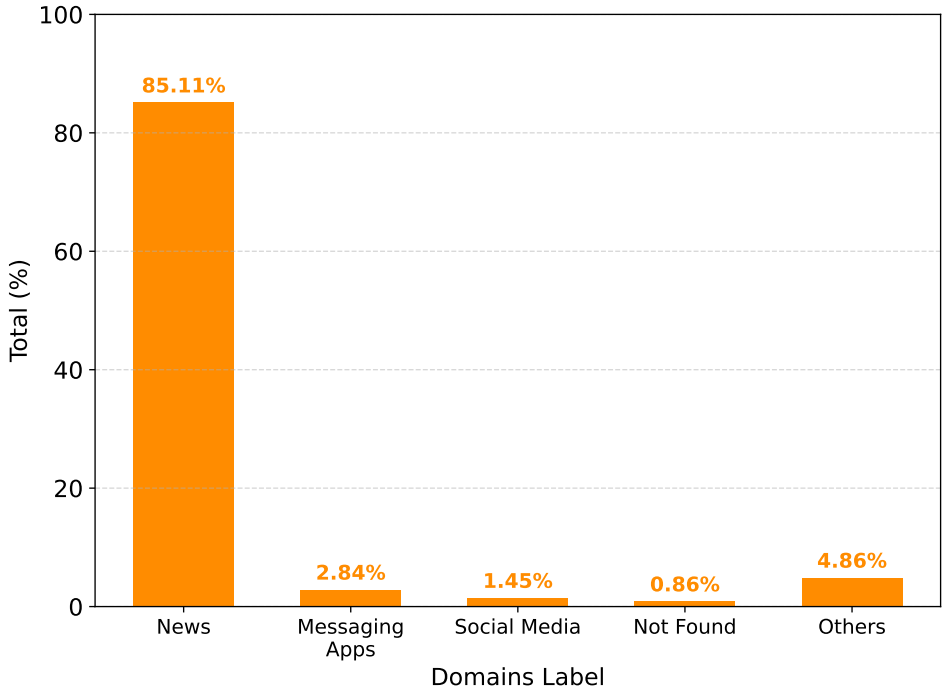
content.

Additionally, we found that 58.9% of coordinated messages were not forwarded. This suggests that most coordinated message sharing is not done through WhatsApp's forwarding mechanism, indicating a deliberate effort to share messages directly, as this mechanism is not widely used. WhatsApp has introduced this feature to combat misinformation and virality by labeling popular forwarded messages with the "forwarded many times" tag [103]. However, our findings suggest that this measure is ineffective in limiting the mass spread of coordinated accounts.

6.3.1 Characterizing Text Messages

Looking at the structure of coordinated text messages, we note an average length of 17.93 words with a standard deviation of 24.19. In contrast, non-coordinated messages are longer, with an average of 24.68 words (±0.4659), but show a much higher standard deviation of 106.07 (±26.74), suggesting significant variation in length. While coordinated messages are generally shorter than the average, the low standard deviation suggests they have a more consistent length than non-coordinated messages, which show greater vari-

Figure 6.6: Category of URLs found in coordinated text messages.



Source: The Author.

ability. This suggests that coordinated messages are not random but contain consistent information to engage users effectively.

6.3.2 Characterizing URLs

Upon closer inspection of the 10,124 coordinated text messages, we observe that 97.31% contain embedded URLs, reflecting the widespread use of hyperlinks by coordinated accounts. These URLs are seamlessly integrated into the text, suggesting that these textual messages not only inform but also direct users to external web resources. Compared with the sample of non-coordinated texts, we find that only 16.91% ($\pm$0.0009) contain links, highlighting a distinct difference between coordinated and non-coordinated messages.

Further analysis of URLs within coordinated messages involved an extraction and manual labeling process. By parsing the coordinated text messages, we found 11,725 links, of which 10,269 were unique. We extracted the domain of each URL from the coordinated text messages, resulting in a subset of 116 unique domains. This indicates that coordinated messages disseminate content from a small number of websites. Here, we characterize the content by undertaking a qualitative analysis to better understand

Domain	Category	Total URLs
pensandodireita*	news	2,403
portaltocanews	news	2,119
redebrasilnews	news	682
gazetabrasil	news	432
whatsapp	messaging apps	293
macajubaacontece	news	291
terrabrasilnoticias*	news	274
brazilnewsinforma	news	263
portalcidade	news	232
direitaonline	news	203

Table 6.1: TOP-10 URLs domains found in coordinated text messages. Domains in bold are sites that employed misinformation strategies during the 2022 electoral campaign, as reported by Aos Fatos Fact Checker

.

the nature of these domains. Initially, we compiled a list of all domains and created a codebook with preliminary codes, refining them iteratively until no further changes were necessary. Our codebook consists of five codes:

- **News**: Websites that host structured news content.
- **Messaging Apps:** Invitations to Telegram and WhatsApp groups/channels.
- **Social Media:** Links to social media platforms such as Facebook, Twitter, Instagram, YouTube, and TikTok.
- **Not Found:** Domains that do not exist or are currently unavailable.
- **Others:** Links that lead to product sales, app stores, or banking services, as well as websites that show characteristics of spam or fraudulent activities.

After building the codebook, we applied it to categorize all domains identified in the coordinated messages. This labeling process allowed us to determine the thematic focus of each shared URL. Figure 6.6 shows the distribution of the coordinated link category. Upon analyzing the results, we observed that 85.11% of all URLs within coordinated messages lead to news websites. Furthermore, messaging apps account for 2.84%, which is particularly interesting because coordinated accounts leverage public platforms to invite others to more exclusive private communities through group invitations. Typically, they share a message that includes an invitation link to other groups.

The significant prevalence of coordinated messages containing links to news websites provides valuable insight into the strategies employed by coordinated users. This suggests a deliberate focus among coordinated users on disseminating news-related content within WhatsApp groups to engage users in propagating specific narratives. By including news links, coordinated users attempt to bring a sense of formality and credibility to their messages. Moreover, including news links serves as a mechanism to drive

traffic to various news websites, thereby expanding the reach and influence of the disseminated information. Notably, websites hosting such coordinated news can generate revenue through advertisements, as shown by the Aos Fatos Fact-Checker.[5]

WhatsApp lacks mechanisms to verify information sources, and this is an ideal scenario in which coordinated users can exploit the platform's ecosystem to disseminate misinformation. This enables coordinated users to leverage WhatsApp as a tool for misinformation, potentially manipulating public opinion. Within political groups, individuals become deeply engaged with particular topics, facilitating the spread of news that aligns with specific narratives. As a result, these coordinated efforts can effectively spread messages that reinforce specific narratives in groups.

Upon analyzing the most frequently shared domains by coordinated users, we found the ten most shared domains in the messages in Table 6.1. Fact-checking agencies revealed that *pensandodireita (1º)* and *terrabrasilnoticias (7º)* domains spread fake news and misinformation. We adopt the definition of misinformation provided by the Aos Fatos fact-checking, which highlights that these websites employed misinformation strategies in their news coverage during the 2022 electoral campaign. According to Aos Fatos, these two websites promote their news across messaging app platforms to attract users to their platforms. These websites typically use multiple ads that eventually engage users who click on these ads, thereby generating additional revenue for the website owner. This discovery sheds light on the deliberate strategies employed to exploit misinformation content for financial gain. Notably, we found that 26% of all links identified in coordinated text messages are from these two misinformation websites, highlighting the role of coordinated accounts in the dissemination of misinformation.

6.3.3 Characterizing Images

Coordinated accounts also use image content to spread narratives, accounting for 15.54% of coordinated messages. Analyzing these images is crucial to understanding the similarities of media types shared in coordinated actions. Observing the ten most shared coordinated messages, we identified that five of these images are related to politics and presidential elections, including promoting candidates or attacking opponents. These images have a different impact on the user's perception, and the visual content can condense more information into a small piece of content. This makes images particularly effective for spreading misinformation, often by presenting events out of context or making old events appear recent. We observed that two of the ten most shared images were mis-

[5] https://aosfatos.org/s/w6gbyzq/

Figure 6.7: TOP-6 coordinated images based on total shares.

(a) (b) (c)

(d) (e) (f)

Source: The Author.

information. The figure 6.8(f), shared 396 times, is the sixth most shared coordinated image. The image is authentic, but the text puts it in the wrong context. Comprova's fact-checking team labeled this image misleading.[6] Another noteworthy example is the fifth most shared coordinated image, as shown in Figure 6.8(e), shared 427 times. This is an authentic picture, but the text puts this image in a different context, saying that one of the members present is a former Supreme Court justice. This image was classified as false by Aos Fatos Fact-checking organization.[7]

Furthermore, two of the analyzed images target the Supreme Federal Court and the electoral process, as shown in Figures 6.8(c) (shared 439 times) and 6.8(e). These two images aim to discredit the judicial decisions and accuse the court of political bias.

[6] https://projetocomprova.com.br/post/re_2B5W8XYjrLpY/
[7] https://aosfatos.org/s/r3i3wti/

This theme was particularly prominent in Brazil, as reflected in the identified topics presented in Table 6.2. A noteworthy aspect of the coordinated images is that some images encourage users to share the message, such as Figure 6.8(b), which has been shared 460 times. Additionally, three of the images are related to credit cards and financial topics. One example is Figure 6.8(d), which depicts a credit card and was shared 430 times.

Takeaways The main takeaways from this section are:

- Coordinated activities focus on news dissemination. 70.24% of coordinated messages consist of text, and 97.31% contain embedded links, with 85.11% of these links leading to political news.
- We found that 26% of all coordinated links are from news misinformation websites.
- Images are a key part of the content shared (15.54%) by coordinated accounts and also include misleading content. Notably, two of the ten most shared images were labeled as misleading content.

6.4 Topic Modeling

Going deeper into the content analysis, we conducted a topic modeling analysis to examine the content of coordinated messages, focusing on identifying their connections with political events. We characterized the topics discussed in coordinated textual messages shared by the coordinated accounts. For this purpose, we employed BERTopic, a topic modeling technique that generates dense clusters to produce interpretable topics. This method uses vector representation (embeddings) and the concept of c-TF-IDF to create coherent topics, preserving key terms in the topic descriptions that enhance the clarity and interpretation [61].

We start by converting coordinated WhatsApp messages into vector embeddings using the PTT5 transformer language model, which was trained on Portuguese Wikipedia data and contains 200 million parameters [20].[8] Next, we apply dimensionality reduction with the Uniform Manifold Approximation and Projection (UMAP) technique, which preserves both local and global structures of the embeddings. Then, we use a hierarchical clustering algorithm (HDBSCAN) to group the vector representations into clusters based on semantic similarities. Finally, we extract topics for each cluster using the Class Term Frequency-Inverse Document Frequency (c-TF-IDF), identifying the most relevant terms

[8] https://huggingface.co/unicamp-dl/ptt5-large-portuguese-vocab

given all documents in a cluster. Then, we refined the approach to balance the number of topics with the size of our dataset. Following the recommendations in the BERTopic documentation, we set the number of topics to 15. To further improve the results, we applied the outlier reduction method to reassign these outlier documents to the appropriate topics. UMAP was configured with ten neighbors, ten components, and a minimum distance of 0.05 for effective dimensionality reduction. Finally, we configured HDBSCAN with a minimum cluster size of 50 and a minimal sample of 50, specifying the minimum number of messages a topic can represent and ensuring that only significant topics are considered for analysis. Additionally, the epsilon parameter was set to 0.3, defining the maximum distance for points to be considered neighbors.

6.4.1 Topic Analyzes

Before analyzing the topics, it is important to understand the political scenario of Brazil during the period analyzed. In 2022, Brazil had a presidential election and, during this period, social networks and messaging apps were flooded with political discussions and campaigns. Furthermore, the analyzed period includes other important events, such as the intense protests marked by widespread fraud allegations about the results and the riots on January 8th. In addition, many protests were made against the decisions of the Supreme Federal Court and its ministers, which became a major topic in the news.

With that in mind, we can observe Table 6.2, which contains the discussed topics identified using the proposed framework. Initially, we can observe many political topics addressing different perspectives on the Brazilian elections. One major focus is the Supreme Court, which became the center of many protests, with several ministers' attacks (Topic 1). Additionally, many messages contain terms related to election fraud (Topic 5), as a large portion of the population refused to accept the results, claiming that it was rigged and asking for military intervention (Topic 15). We also observe discussions about journalists (Topic 6), political candidates (Topic 8), and political debate repercussions (Topic 10). Also, some topics reflect government arguments promoting the candidates' achievements (Topic 7), which relate to fuel and price reductions. Coordinated messages also include government assistance and subsidies for low-income individuals (Topic 14). Initially, we observed that the coordinated messages were related to the Brazilian political scenario, which gave us an idea of the interest of coordinated accounts in reinforcing specific narratives.

Furthermore, we identified topics unrelated to politics, such as messages about credit cards, bank loans, and financial resources (Topics 2 and 13). There is also content

Id	Label	Topic Terms
1	**Supreme Court of Justice - stf (42.9%)**	moraes, stf, pt, federal, minister, tse, round, whatsapp, police, pt supporter
2	**Credit Card (15%)**	thousand, card, credit, aid, cash, millions, bank, vacancies, request, receive
3	**Videos (8%)**	video, audio, activate, screen, husband, caught, wife, lover, videos, shows
4	**Social Networks (5.1%)**	telegram, whatsapp, twitter, network, social, facebook, instagram, follow, forget, follow us
5	**Election Fraud (4.6%)**	electoral court, electoral, crime, crimes, propaganda, fraud, operation, federal police, ballot boxes, elections
6	**Journalist Commentators (2.9%)**	shorts, amanda, klein, *toma, invertida, lapada*, guga, noblat, come through, live
7	**Economy and Fuel Price (3.5%)**	petrobras, price, gasoline, pf, seize, gas, reduction, tons, fuels, federal highway police (prf)
8	**Candidates (3.2%)**	candidate, candidates, presidency, candidacy, health, agenda, curses, covid-19, education, childish
9	**Health (2.4%)**	symptoms, cancer, know, hospital, disease, main, heart attack, treatment, signs, warning
10	**Political Debate (2.4%)**	debate, tv, criticize, band tv, knight, diego, globo tv, democracy, criticism, sbt tv
11	**Accident News (0.33%)**	death, dead, leaves, accident, found, dies, serious, deaths
12	**Pictures (1.8%)**	pictures, picture, images, body, globo tv, bikini, cameras, camera, happens, attention
13	**Money (1.8%)**	loan, aid, entrepreneurs, companies, payroll, beneficiaries, moraes, name, businessman, bank caixa
14	**Government Assistance (1.8%)**	family, families, scholarship, aid, receive, low, income, program, own
15	**Military (0.13%)**	military, military, police, defense, security, institutional security office, minister fachin, civil, organization, superior

Table 6.2: Discussion topics found in coordinated text messages. The topic terms are translated as the original in Brazilian Portuguese. The percentages in the labels represent the proportion of coordinated text messages for each topic.

that involves shared images, often featuring sexist themes (Topic 12). Videos are also popular, as seen in Topic 3, where many messages contain links to external websites hosting the videos. We also found health-related content that spreads tips on disease symptoms (Topic 9) and news about violence and accidents (Topic 11). As observed in domain analysis, many websites focus on increasing traffic and want to promote their content, often about curiosity or facts that intrigue people to know more, characteristics observed in these two identified topics.

The identified topics highlight the main discourses and narratives of the Brazilian political landscape during the period. WhatsApp plays a central role in Brazil's communication ecosystem, widely used for debates, gathering information, and tracking the repercussions of major events. Although WhatsApp is typically expected to be used to react and discuss real-world events, these coordinated actions suggest that, in some cases, it can be used as a strategy to organize, motivate, mobilize, and shape political narratives.

To better understand coordinated messages and their motivations, we focus on two key topics: the Supreme Federal Court (Topic 1) and Election Fraud (Topic 6). These topics were chosen because they were central to the Brazilian political discourse during the period covered by our dataset. The election is the most significant event, making these topics particularly relevant for analyzing how coordinated messages were used to influence the discussions. The Supreme Federal Court (Topic 1) became a focal point of political debates during the election, which was marked by widespread protests and many allegations of election fraud (Topic 6), making it a critical subject in public discourse.

6.4.2 Case Study 1

In this case, we identified a coordinated attack targeting the Brazilian Supreme Federal Court (STF), specifically aimed at doxxing the locations of the ministers. During the analyzed period captured by our dataset, the members of the Supreme Court became focal points of intense online discussions, which can be observed by Topic 1. To better understand this content, we analyzed the messages related to this topic and selected the most widely shared coordinated message about the ministers, which is shown as follows:

Message

"We have just discovered the hotel where the ministers of the Supreme Federal Court are staying in New York, please forward this to all Brazilians in the USA. xxxW xxth St, New York, NY xxxxx, United States" (translated and anonymized to not show address).

This message was shared 144 times within the entire dataset, with 29.2% of those shares coming from coordinated activity. The message reached 102 WhatsApp groups (7% of the total dataset), posted by 42 coordinated accounts.

Context. In November, following the Brazilian presidential election, there were tensions surrounding the decisions of the Supreme Court. When the ministers traveled to New York for a conference on November 13th, their hotel location was maliciously doxxed online, inciting an attack. This led to a flood of messages on WhatsApp, encouraging people to gather and confront the ministers.

Analysing the Impact. The dissemination of this message had a significant real-world impact. Protesters quickly mobilized to the location, gathering outside the hotel on the night of November 13. The ministers faced harassment and confrontations when entering and leaving the hotel.[9] The rapid spread of this information was crucial in organizing these protests, demonstrating how coordinated actions on WhatsApp can escalate from the digital ecosystem to a physical response. The incident highlights the power of doxxing, coordinated by some users, to endanger public figures by inciting hostile actions within hours.

6.4.3 Case Study 2

Here, we observed the Electoral Fraud (Topic 5). In this case, we analyzed the topic of Electoral Fraud. We searched for coordinated messages using the terms "fraud" and "ballot boxes". From the top three most relevant messages based on the total number of shares, we found the following message:

Message

"Our president said ALL PEACEFUL PROTESTS ARE WELCOME!!!! COME ON MY PEOPLE!! WE WILL NOT BACK DOWN! ALL THE RIGHT-WING IS GOING TO TAKE TO THE STREETS - THE ARMED FORCES ARE JUST WAITING TO REACH THE NUMBER TO HAVE THE NATIONAL AND INTERNATIONAL QUORUM THAT IS THE MASS OF THE POPULATION IN THE STREETS TO MEET THE DEMANDS OF THE PEOPLE" (translated).

[9] https://www.cnnbrasil.com.br/politica/manifestantes-hostilizam-ministros-do-stf-na-porta-de-h

This message was shared 83 times during the analyzed period, and 24% of these shares were in coordinated activities. It reached 74 different groups, and two coordinated accounts were involved in the coordinated actions of this message.

Context. After the election, former President Bolsonaro made his first public statement about the election. After that, messages like this one began to be shared, claiming that his pronouncement contained a subliminal message encouraging people to go to military posts and demand military intervention due to the alleged fraudulent election result.

Analysing the Impact. This coordinated message was widely shared after Bolsonaro's speech. After that, the protests intensified.[10] Many people took to the streets in front of military posts, demanding military intervention and alleging fraud in the polls. This coordinated message reinforced and motivated users to continue protesting and taking specific actions that had a real impact on society. This kind of message needs to be spread quickly, and the rapid coordinated actions work perfectly in this context, reaching more people quickly.

By examining these coordinated examples, we can reinforce that WhatsApp is an extremely politically relevant tool in Brazil, and coordinated activities can have a much greater influence, expanding discourse and reinforcing narratives. In our context, we observed that coordinated actions impacted orchestrating real-world events, such as protests and specific mobilizations. These coordinated actions aim to reach as many people as possible quickly, which is evident in both cases analyzed.

Takeaways. The main takeaways from this section are:

- The proposed topic analysis reveals that key Brazilian political events are highlighted in coordinated messages, particularly those that raise suspicions about electoral fraud and call for military intervention.
- We found evidence that the events discussed in the case studies (i.e., the mobilization of people against Supreme Court justices and protests over the election results) were also driven by coordinated accounts in WhatsApp groups.

[10] https://www.reuters.com/world/americas/bolsonaro-backers-call-brazil-military-intervene-after

6.5 Summary

This chapter provides valuable insights into the coordinated activities driving message propagation within WhatsApp. By examining 13M in public political groups from July 2022 to January 2023, we found a significant prevalence of coordinated accounts in disseminating messages on the platform. Our analysis reveals the presence of 1.5K coordinated accounts that work simultaneously to disseminate messages across multiple groups. These coordinated activities are focused on spreading news text messages that can easily shared across various groups. Our investigation showed that 26% of the links shared are from misinformation websites. This strategic news dissemination aims to engage audiences and promote specific political viewpoints. Furthermore, we observed that coordinated actions had a significant impact, as they were used to previously orchestrate protests and important political actions in the Brazilian political scenario.

Overall, this chapter sheds light on a frequent but little-explored phenomenon on WhatsApp. While the influence of misinformation and message apps on society is well recognized, the influence of coordinated activities is quite new. Our research reveals compelling evidence of coordinated efforts to disseminate misinformation on the platform and mobilize people to specific events. Importantly, even though we can not access the entire WhatsApp network, we identified many coordinated accounts that are working on spreading messages by public groups. This suggests that the problem is likely to be even larger than observed. Even with WhatsApp recognizing the existence of coordinated campaigns in the last Brazilian elections [99] and limiting forwarding per user to control virality [102], it was not enough to solve the problem, especially because coordinated accounts do not frequently use forwarding tool to spread their content. By observing the real impact employed by coordinated users, it is necessary to more notable strategies to mitigate the problem of controlling the influence of coordinated accounts. In 2018, WhatsApp already banned hundreds of thousands of accounts detected as spammers in Brazil.[11] Banning is a strategy that temporarily mitigates the problem, but it needs to be aligned with other control policies to keep the WhatsApp environment healthier. Even because new features were introduced facilitating massive spreading, such as increasing the number of participants per group and creating communities.

The methods to combat coordination misinformation should consider simultaneous posting patterns to mitigate coordination actions between multiple groups or accounts, since simultaneous posting activity is an important factor in the effectiveness of coordinated campaigns, such as protests or organized attacks. In this regard, effectively addressing coordinated misinformation is essential to foster collaboration between platforms

[11] https://wapo.st/3ZudkSD

and government authorities. This collaboration is particularly critical in high-impact contexts such as elections, where misinformation and coordinated accounts can directly and harmfully impact society. Increasing transparency from platforms is crucial, which should not only involve reporting on coordinated campaigns but also include measures to restrict the volume of messages during critical periods, improve content moderation algorithms, and ensure the detection and removal of harmful and inauthentic activity.

Chapter 7

Summary of Results and Next Steps

In this chapter, we provide an overview of the thesis, summarizing the key findings to date, and presenting the planned schedule to conclude this thesis.

7.1 General Outline

Given the open questions that guide this thesis, we have made partial progress in addressing the research goals presented in Chapter 1. In this section, we provide a general overview of the work completed so far, describing the methodology, and key findings, and outline the planned steps to finalize the thesis.

RG1 – Understanding attacks and hostile interactions in public groups.

In this research goal, we investigated the dynamics of WhatsApp groups and proposed a methodology that identified two distinct types of disruptive attacks. Particularly, while flooding attacks aim to disseminate many messages within a short period, hijacking attacks aim to take control of the group and wipe them out. After identifying the attacks, we analyze the occurrences and their impact. We identified that flooding attacks are not rare and groups are the recipients of multiple flooding attacks, even within the same day, which likely highlights the lack of effective tools that assist the group moderators. Subsequently, the messages used in these attacks were characterized. The analysis revealed a significant prevalence of stickers, as well as the presence of harmful content, including offensive, discriminatory, and violent material. This work contributes to understanding the negative aspects of WhatsApp group interactions, particularly hostile intergroup interactions across the political spectrum.

Publication: This work resulted in a published paper [74]:

- *Kansaon, D., de Freitas Melo, P., Zannettou, S., Feldmann, A., Benevenuto, F. (2024, May). Strategies and Attacks of Digital Militias in WhatsApp Political Groups. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 813-825).*

RG2 – Understanding stickers as a new form of spreading harmful content.

In this research goal, we explored how stickers have been appropriated as a mechanism for spreading harmful content within WhatsApp political groups. We proposed a methodology to detect sticker interactions and investigated their usage patterns. Our analysis revealed that stickers are not only widely adopted in political discourse but are also strategically used in coordinated attacks. In particular, we observed their role in flooding attacks, where stickers are employed to overwhelm conversations and disseminate offensive or misleading content in public groups, often without any form of moderation. Furthermore, we conducted a qualitative analysis of the most frequently shared stickers and identified the presence of abusive content, including offensive, provocative, and politically charged imagery. These findings demonstrate the structural uniqueness of sticker-based communication and reveal distinct usage patterns, including a darker side of sticker usage. Overall, our results show that stickers play a significant role in shaping political discourse on WhatsApp and underscore the importance of considering multimodal content in efforts to moderate harmful behavior on messaging platforms.

Publication: This work resulted in a published paper [34]:

- *de Freitas Melo, P., Kansaon, D., Couto, J. M., Reis, J. C., Benevenuto, F. (2025, June). A Sticker is Worth a Thousand Words: Characterizing the Use and Abuse of Stickers on WhatsApp Political Groups in Brazil. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 19, pp. 1210-1223).*

RG3 – Identifying and characterizing coordinated strategies employed for message spreading.

In this context, we proposed a rapid coordination network model designed to identify coordinated efforts that enable the quick spread of messages. The first step in this process is to detect coordinated users. To achieve this, we model all message interactions between groups and observe users who share similar messages at the same time. We analyzed 13M in public political groups from July 2022 to January 2023. Our findings revealed a significant prevalence of coordinated accounts in disseminating messages on the platform. Specifically, we identified 1.5K coordinated accounts that work simultaneously to disseminate messages across multiple groups. Our analysis showed that 26% of the links shared are from misinformation websites. Furthermore, we applied the BERTopic strategy to extract the topics of coordinated messages. We observed that coordinated messages addressed specific relevant political topics and were used to orchestrate protests and important political actions in the Brazilian political scenario.

Publication: This work resulted in a published paper [73]:

- *Kansaon, D., de Freitas Melo, P., Zannettou, S., Benevenuto, F. (2025, June). From Fake News to Real Protests: WhatsApp's Role in Brazilian Political Coordi-*

Table 7.1: Planned schedule for the conclusion of the project

	Month						
Activity	**Sep 25**	**Oct 25**	**Nov 25**	**Dec 25**	**Jan 25**	**Feb 25**	**Mar 25**
Create a dataset labeled with harmful text	•	•					
Qualitative and quantitative content analysis		•	•	•			
Survey of state-of-the-art approaches		•	•	•	•	•	
Thesis writing and refinement		•	•	•	•	•	
Final defense and submission							•

nation. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 19, pp. 1007-1020).

RG4 – Identifying subtle and harmful text-based discourse.

This research objective is still ongoing, and we are working to achieve significant progress. By the end of the thesis, we hope to better understand the WhatsApp political text message dataset, seeking to uncover recurring linguistic and narrative patterns and assess their role in the broader dynamics of abusive and manipulative discourse on the platform. Our focus is to analyze messages shared in groups to identify harmful textual content. Here, we concentrate on developing methods to identify linguistic and narrative patterns of harmful text. Our initial efforts have focused on analyzing stickers, revealing instances of harmful material such as Nazi, racist, and various forms of abusive content. Currently, we are shifting our focus to text-based content, specifically aiming to identify those that are particularly prevalent in political discussions.

7.2 Planned Scheduled

Table 7.1 presents the timeline for the final phase of this thesis. The planned activities include: (i) the construction of a labeled dataset with harmful text; (ii) qualitative and quantitative content analysis of the harmful messages; and (iii) the development of a systematic survey aimed at mapping the research area, consolidating the current state of the art, and identifying existing research gaps. Although many studies have addressed related topics, no comprehensive effort has been made to map these contributions and analyze their thematic scope. This survey will be conducted throughout the remainder of the thesis and will be incorporated into the related works section.

References

[1] Fatimah Alkomah and Xiaogang Ma. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273, 2022.

[2] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L. Williams. "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web*, 13(3), July 2019.

[3] Abdullah Alrhmoun, Charlie Winter, and János Kertész. Automating terror: The role and impact of telegram bots in the islamic state's online ecosystem. *Terrorism and political violence*, 36(4):409–424, 2024.

[4] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. A deep dive into multilingual hate speech classification. In *Machine learning and knowledge discovery in databases. Applied data science and demo track: European conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, proceedings, Part V*, pages 423–439. Springer, 2021.

[5] Lorenzo Alvisi, Serena Tardelli, and Maurizio Tesconi. Unraveling the italian and english telegram conspiracy spheres through message forwarding. In Luca Maria Aiello, Tanmoy Chakraborty, and Sabrina Gaito, editors, *Social Networks Analysis and Mining*, pages 204–213, Cham, 2025. Springer Nature Switzerland.

[6] Chinmayi Arun. On WhatsApp, Rumours, and Lynchings. *Economic & Political Weekly*, 54(6):30–35, 2019.

[7] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, pages 1–12, 2024.

[8] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[9] Morten Bay. The Political Life of GIFs: The Common Emotional Displays That Let Us Disagree More Consummately on Social Media. *InVisible Culture*, (34), 2022.

[10] Fabrício Benevenuto and Philipe Melo. Misinformation campaigns through whatsapp and telegram in presidential elections in brazil. *Communications of the ACM*, 67(8):72–77, aug 2024.

[11] Fabrício Benevenuto and Philipe Melo. Misinformation campaigns through whatsapp and telegram in presidential elections in brazil. *Communications of the ACM*, 67(8):72–77, 2024.

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[13] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[14] Keith Burghardt, Ashwin Rao, Georgios Chochlakis, Baruah Sabyasachee, Siyi Guo, Zihao He, Andrew Rojecki, Shrikanth Narayanan, and Kristina Lerman. Sociolinguistic characteristics of coordinated inauthentic accounts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 164–176, 2024.

[15] Pete Burnap and Matthew L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[16] Victor S. Bursztyn and Larry Birnbaum. Thousands of small, constant rallies: A large-scale analysis of partisan whatsapp groups. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 484–488, 2019.

[17] Victor S. Bursztyn and Larry Birnbaum. Thousands of small, constant rallies: a large-scale analysis of partisan whatsapp groups. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 484–488, New York, NY, USA, 2020. Association for Computing Machinery.

[18] Antoine Buyse. Words of violence: "fear speech," or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, 36(4):779–797, 2014.

[19] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T. Marques-Neto, and Virgílio Almeida. Characterizing attention cascades in whatsapp groups. In *Proceedings of the 10th ACM Conference on Web Science*,

WebSci '19, page 27–36, New York, NY, USA, 2019. Association for Computing Machinery.

[20] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

[21] Tatiana Cesso. Meta Reveals Brazilians are Number 1 in Voice Messaging. https://brazilcore.com/meta-reveals-brazilians-number-1-voice-messaging/, 2023. Accessed: 2024-04-03.

[22] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 13–22, New York, NY, USA, 2017. Association for Computing Machinery.

[23] Nic Cheeseman, Jonathan Fisher, Idayat Hassan, and Jamie Hitchen. Social media disruption: Nigeria's whatsapp politics. *Journal of Democracy*, 31(3):145–159, 2020.

[24] Di (Laura) Chen, Dustin Freeman, and Ravin Balakrishnan. Integrating multimedia tools to enrich interactions in live streaming for language learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[25] Federico Cinus, Marco Minici, Luca Luceri, and Emilio Ferrara. Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 u.s. election. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 541–559, New York, NY, USA, 2025. Association for Computing Machinery.

[26] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[27] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):89–96, Aug. 2021.

[28] Marlen Couto. De tática para furar bolha a termômetro emocional do eleitor, figurinhas de WhatsApp invadem disputa presidencial (in Portuguese). http://bit.ly/4fyK1UW, 2022. Accessed: 2025-04-03.

[29] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71, 2015.

[30] Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. 13(2), April 2019.

[31] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11, 03 2017.

[32] Patrick Davison. The language of internet memes. *The social media reader*, pages 120–134, 2012.

[33] Philipe de Freitas Melo. *Activism and misinformation on WhatsApp: measurement, analysis, and countermeasures*. Phd thesis, Dept. of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, 2022.

[34] Philipe de Freitas Melo, Daniel Kansaon, João MM Couto, Julio CS Reis, and Fabricio Benevenuto. A sticker is worth a thousand words: Characterizing the use and abuse of stickers on whatsapp political groups in brazil. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1210–1223, 2025.

[35] Dingsheng Deng. Dbscan clustering algorithm based on density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pages 949–953, 2020.

[36] Statista Research Department. Number of social media users worldwide from 2010 to 2024 (in billions). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2024. Accessed: 2025-03-14.

[37] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. COMPA: Detecting Compromised Accounts on Social Networks. In *Proceedings of the 20th Symposium on Network and Distributed System Security*, San Diego, CA, 2013.

[38] Brian Ellsworth and Rodrigo Viga Gaier. Bolsonaro backers call on Brazil military to intervene after Lula victory. http://bit.ly/3UYd8HJ, Out 2022. Accessed: 2024-12-02.

[39] Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. Misleading repurposing on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):209–220, Jun. 2023.

[40] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[41] Carlos Elías and Daniel Catalan-Matamoros. Coronavirus in spain: Fear of 'official' fake news boosts whatsapp and alternative sources. *Media and Communication*, 8(2):462, 2020.

[42] Estadão. 8 de janeiro: O mês dos ataques golpistas e invasão em brasília – o que se sabe. https://www.estadao.com.br/politica/8-janeiro-mes-ataques-golpistas-invasao-brasilia-o-que-se-sabe/, 2023. Accessed: 2025-03-14.

[43] EuroPol. Operation Chemosh: how encrypted chat groups exchanged emoji 'stickers' of child sexual abuse. http://bit.ly/47sVJOU, 2019. Accessed: 2024-04-03.

[44] Facebook. How to Create and Upload Subscriber Stickers on Facebook. https://www.facebook.com/business/help/767160797072449, 2022. Accessed: 2025-04-03.

[45] Senado Federal. Projeto de lei nº [número do projeto] - dispõe sobre a proibição do uso de robôs (bots) em redes sociais. https://www25.senado.leg.br/web/atividade/materias/-/materia/141944, 2020. Accessed on 2025-02-21.

[46] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[47] X (formerly Twitter). Hateful conduct policy. https://help.x.com/en/rules-and-policies/hateful-conduct-policy, 2023. Accessed: 2025-02-24.

[48] Antônio Diogo Forte Martins, Lucas Cabral, Pedro Jorge Chaves Mourão, José Maria Monteiro, and Javam Machado. Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *International Conference on Applications of Natural Language to Information Systems*, pages 199–206. Springer, 2021.

[49] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July 2018.

[50] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and

Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[51] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293, 2020.

[52] Kiran Garimella, Princessa Cintaqia, Juan José Rojas-Constain, Bharat Kumar Nayak, and Aditya Vashistha. Global patterns of viral content on whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):586–601, Jun. 2025.

[53] Kiran Garimella and Dean Eckles. Images and misinformation in political groups: Evidence from whatsapp in india. *Harvard Kennedy School Misinformation Review*, 2020.

[54] Kiran Garimella and Gareth Tyson. Whatapp doc? a first look at whatsapp public group data. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[55] Jing Ge. The anatomy of memetic stickers: An analysis of sticker competition on chinese social media. In *Proc. of the ICWSM Workshop*, volume 10, 2020.

[56] Patrick Gerard, Svitlana Volkova, Louis Penafiel, Kristina Lerman, and Tim Weninger. Modeling information narrative evolution on telegram during the russia-ukraine war. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):602–614, Jun. 2025.

[57] Patrick Gerard, Tim Weninger, and Kristina Lerman. Fear and loathing on the frontline: Decoding the language of othering by russia-ukraine war bloggers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 615–635, 2025.

[58] Fabio Giglietto, Nicola Righetti, Luca Rossi, and Giada Marino. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 italian elections. *Information, Communication & Society*, 23(6):867–891, 2020.

[59] Christian Grimme, Dennis Assenmacher, and Lena Adam. Changing perspectives: Is it sufficient to detect social bots? In *International conference on social computing and social media*, pages 445–461. Springer, 2018.

[60] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[61] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[62] Samuel Guimaraes, Gabriel Kakizaki, Philipe Melo, Márcio Silva, Fabricio Murai, Julio CS Reis, and Fabrício Benevenuto. Anatomy of hate speech datasets: Composition analysis and cross-dataset classification. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pages 1–11, 2023.

[63] Hans WA Hanley and Zakir Durumeric. Partial mobilization: Tracking multilingual information flows amongst russian media outlets and telegram. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 528–541, 2024.

[64] Hatebase. Hatebase: The world's largest structured repository of multilingual hate speech. https://hatebase.org/, 2025. Accessed: 2025-02-26.

[65] Mohamad Hoseini, Philipe Melo, Fabricio Benevenuto, Anja Feldmann, and Savvas Zannettou. On the globalization of the qanon conspiracy theory through telegram. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 75–85, New York, NY, USA, 2023. Association for Computing Machinery.

[66] Mohamad Hoseini, Philipe Melo, Manoel Junior, Fabrício Benevenuto, Balakrishnan Chandrasekaran, Anja Feldmann, and Savvas Zannettou. Demystifying the messaging platforms' ecosystem through the lens of twitter. In *IMC*, 2020.

[67] Bilal Hussain, Qinghe Du, Bo Sun, and Zhiqiang Han. Deep learning-based ddos-attack detection for cyber–physical system over 5g network. *IEEE Transactions on Industrial Informatics*, 2021.

[68] Vincenzo Imperati, Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Francesco Sassi. The conspiracy money machine: Uncovering telegram's conspiracy channels and their profit model. *arXiv preprint arXiv:2310.15977*, 2023.

[69] R. Tallal Javed, Mirza Elaaf Shuja, Muhammad Usama, Junaid Qadir, Waleed Iqbal, Gareth Tyson, Ignacio Castro, and Kiran Garimella. A first look at covid-19 messages on whatsapp in pakistan. In *Proceedings of the 12th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '20, page 118–125. IEEE Press, 2021.

[70] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[71] Google Jigsaw. Perspective api. https://perspectiveapi.com/, 2025. Accessed: 2025-02-24.

[72] Manoel Júnior, Philipe Melo, Daniel Kansaon, Vitor Mafra, Kaio Sa, and Fabricio Benevenuto. Telegram monitor: Monitoring brazilian political groups and channels on telegram. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 228–231, 2022.

[73] Daniel Kansaon, Philipe de Freitas Melo, Savvas Zannettou, and Fabricio Benevenuto. From fake news to real protests: Whatsapp's role in brazilian political coordination. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1007–1020, 2025.

[74] Daniel Kansaon, Philipe F. Melo, Savvas Zannettou, Anja Feldmann, and Fabrício Benevenuto. Strategies and attacks of digital militias in whatsapp political groups. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 813–825, 2024.

[75] Daniel Pimentel Kansaon, Philipe De Freitas Melo, and Fabrício Benevenuto. "click here to join": A large-scale analysis of topics discussed by brazilian public groups on whatsapp. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 55–65, New York, NY, USA, 2022. Association for Computing Machinery.

[76] Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, and Scott A Hale. Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 indian general election on whatsapp. *(HKS) Misinformation Review*, 3(1), 2022.

[77] Mateusz Kazimierczak, Thanyathorn Thanapattheerakul, and Jonathan H. Chan. Enhancing security in whatsapp: A system for detecting malicious and inappropriate content. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT '23, page 274–281, New York, NY, USA, 2023. Association for Computing Machinery.

[78] Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political communication*, 37(2):256–280, 2020.

[79] Baris Kirdemir and Oluwaseyi Adeliyi. Towards characterizing coordinated inauthentic behaviors on youtube. In *The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022).*, 2023.

[80] Adam Klein. *Fanaticism, racism, and rage online: Corrupting the digital sphere*. Springer, 2017.

[81] Samantha Klier and Harald Baier. To possess or not to possess-whatsapp for android revisited with a focus on stickers. In *Nordic Conference on Secure IT Systems*, pages 281–303. Springer, 2023.

[82] Ian Kloo, Iain J. Cruickshank, and Kathleen M. Carley. A cross-platform topic analysis of the nazi narrative on twitter and telegram during the 2022 russian invasion of ukraine. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):839–850, May 2024.

[83] Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. Predictive embeddings for hate speech detection on Twitter. In Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[84] Danang Arif Kurniawan. Criminal penalties for individuals who create stickers with other people's faces. *Gema Wiralodra*, 15(1):145–150, Jan. 2024.

[85] Febbie Austina Kwanda and Trisha T. C. Lin. Fake news practices in indonesian newsrooms during and after the palu earthquake: a hierarchy-of-influences approach. *iCS*, 23(6):849–866, 2020.

[86] Joon Young Lee, Nahi Hong, Soomin Kim, Jonghwan Oh, and Joonhwan Lee. Smiley face: Why we use emoticon stickers in mobile messaging. In *Proc. of the Int'l Conf. on Human-Computer Interac. with Mobile Devices and Services Adjunct*, page 760–766, 2016.

[87] Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Dissecting the meme magic: Understanding indicators of virality in image memes. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), 2021.

[88] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

[89] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. In *Companion Proceedings of The 2019 World Wide Web Con-*

ference, WWW '19, page 1013–1019, New York, NY, USA, 2019. Association for Computing Machinery.

[90] Golshan Madraki, Isabella Grasso, Jacqueline M. Otala, Yu Liu, and Jeanna Matthews. Characterizing and comparing covid-19 misinformation across languages, countries and platforms. In *Companion proceedings of the web conference 2021*, pages 213–223, 2021.

[91] Jay Mahadeokar and Gerry Pesavento. Open Sourcing a Deep Learning Solution for Detecting NSFW Images. https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for, 2016. Accessed: 2025-04-03.

[92] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17, 2019.

[93] Lorenzo Mannocci, Michele Mazza, Anna Monreale, Maurizio Tesconi, and Stefano Cresci. Detection and characterization of coordinated online behavior: A survey. *arXiv preprint arXiv:2408.01257*, 2024.

[94] Alexandre Maros, Jussara Almeida, Fabrício Benevenuto, and Marisa Vasconcelos. Analyzing the use of audio messages in whatsapp groups. In *Proceedings of The Web Conference 2020*, WWW '20, page 3005–3011, New York, NY, USA, 2020. Association for Computing Machinery.

[95] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Association for Computing Machinery*, WebSci '19, page 173–182, New York, NY, USA, 2019.

[96] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[97] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019.

[98] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[99] Patrícia Campos Mello. WhatsApp admite envio maciço ilegal de mensagens nas eleições de 2018. https://www1.folha.uol.com.br/internacional/en/brazil/2019/10/whatsapp-admits-to-illegal-mass-messaging-in-brazils-2018.shtml, Out 2019. Accessed: 2024-09-13.

[100] Patrícia Campos Mello and Renata Galf. TSE dá ordens em série para derrubar grupos golpistas que se multiplicam nas plataformas. http://bit.ly/4lpooYA, Sep 2022. Accessed: 2025-03-13.

[101] Philipe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. Whatsapp monitor: A fact-checking system for whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):676–677, Jul. 2019.

[102] Philipe Melo, Carolina Vieira, Kiran Garimella, Pedro O.S. Vaz de Melo, and Fabricio Benevenuto. Can whatsapp counter misinformation by limiting message forwarding? In *Proceedings of the International Conference on Complex Networks and their Applications*, 2019.

[103] Philipe F. Melo, Mohamad Hoseini, Savvas Zannettou, and Fabricio Benevenuto. Don't break the chain: Measuring message forwarding on whatsapp. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18 of *ICWSM'24*, pages 1054–1067, 2024.

[104] Meta. Community standards enforcement report on hate speech (instagram). https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/instagram/, 2025. Accessed: 2025-02-24.

[105] Andrés Moreno, Philip Garrison, and Karthik Bhat. Whatsapp for monitoring and response during critical events: Aggie in the ghana 2016 election. In *14th International Conference on Information Systems for Crisis Response and Management*, pages 645–655, 2017.

[106] N Newman, R Fletcher, CT Robertson, A Ross Arguedas, and RK Nielsen. Reuters institute digital news report 2024. Technical report, 2024.

[107] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. *Reuters Institute Digital News Report 2021*. Reuters Institute for the Study of Journalism, 2021.

[108] Facebook Newsroom. Inside feed: Coordinated inauthentic behavior. https://about.fb.com/news/2018/12/

inside-feed-coordinated-inauthentic-behavior/, December 2018. Accessed on 2025-02-21.

[109] Taberez Ahmed Neyazi, Aaron Yi Kai Ng, Ozan Kuru, and Burhanuddin Muhtadi. Who gets exposed to political misinformation in a hybrid media environment? the case of the 2019 indonesian election. *Social media+ society*, 8(3):20563051221122792, 2022.

[110] Lynnette Hui Xian Ng, Ian Kloo, Samantha Clark, and Kathleen M Carley. An exploratory analysis of covid bot vs human disinformation dissemination stemming from the disinformation dozen on telegram. *Journal of Computational Social Science*, 7(1):695–720, 2024.

[111] Jack Nicas and André Spigariol. Bolsonaro Supporters Lay Siege to Brazil's Capital. https://nyti.ms/49m15sB, 2023. Accessed: 2023-05-15.

[112] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Coordinated behavior on social media in 2019 uk general election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 443–454, 2021.

[113] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[114] Gabriel Peres Nobre, Carlos Henrique Gomes Ferreira, and Jussara Marques Almeida. Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 252–266, 2020.

[115] Gabriel Peres Nobre, Carlos HG Ferreira, and Jussara M Almeida. A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. *Information Processing & Management*, 59(1):102757, 2022.

[116] Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, and Silvia Giordano. Misinformation and polarization around covid-19 vaccines in france, germany, and italy. In *Proceedings of the 16th ACM web science conference*, pages 119–128, 2024.

[117] Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. Toxic bias: Perspective api misreads german as more toxic. In *Proceedings of the international AAAI conference on web and social media*, volume 19, pages 1346–1357, 2025.

[118] Diogo Pacheco, Alessandro Flammini, and Filippo Menczer. Unveiling coordinated groups behind white helmets disinformation. In *Companion proceedings of the web conference 2020*, pages 611–616, 2020.

[119] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering coordinated networks on social media: methods and case studies. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 455–466, 2021.

[120] Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*, 2020.

[121] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

[122] Julio Reis, Philipe de Freitas Melo, Kiran Garimella, and Fabrício Benevenuto. Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *arXiv preprint arXiv:2006.02471*, 2020.

[123] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

[124] Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and India Elections. *ICWSM*, 14(1):903–908, May 2020.

[125] Gustavo Resende, Philipe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara M. Almeida, and Fabrício Benevenuto. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In *WebSci*, pages 225–234, 2019.

[126] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The Web Conference*, page 818–828. ACM, 2019.

[127] Alison Ribeiro and Nádia Silva. Inf-hateval at semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 420–425, 2019.

[128] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[129] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.

[130] Jon Russell. Stickers: From Japanese craze to global mobile messaging phenomenon. https://thenextweb.com/news/stickers, 2013. Accessed: 2025-04-03.

[131] Jon Russell. Chat app Line makes over $270 million a year from selling stickers. https://techcrunch.com/2016/06/13/chat-app-line-makes-over-270-million-a-year-from-selling-stickers/, 2016. Accessed: 2025-04-03.

[132] Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120, 2023.

[133] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. In *Proceedings of the Web conference 2021*, pages 1110–1121, 2021.

[134] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[135] Elyse Samuels. How misinformation on WhatsApp led to a mob killing in India. https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india/, Feb 2020. Accessed: 2025-02-26.

[136] Karsten Schmehl. WhatsApp Has Become A Hotbed For Spreading Nazi Propaganda In Germany. https://www.buzzfeednews.com/article/karstenschmehl/whatsapp-groups-nazi-symbol-stickers-germany, 2019. Accessed: 2024-04-03.

[137] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro O. Vaz de Melo, Oana Goga, and Fabrício Benevenuto. Facebook ads monitor: An independent

auditing system for political ads on facebook. In *Proceedings of The Web Conference (WWW'20)*, 2020.

[138] Ivan Smirnov, Camelia Oprea, and Markus Strohmaier. Toxic comments are associated with reduced activity of volunteer editors on wikipedia. *PNAS nexus*, 2(12):pgad385, 2023.

[139] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–26, 2019.

[140] Statista. Most popular social networks worldwide as of april 2024, by number of monthly active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, Apr 2024. Accessed: 2025-03-12.

[141] Elisabeth Steffen. More than memes: A multimodal topic modeling approach to conspiracy theories on telegram. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1831–1844, Jun. 2025.

[142] Neil Thompson. *Anti-discriminatory practice: Equality, diversity and social justice*. Bloomsbury Publishing, 2020.

[143] Tribunal Superior Eleitoral. TSE e WhatsApp lançam pacote de figurinhas para as Eleições 2022. https://www.tse.jus.br/comunicacao/noticias/2022/Setembro/tse-e-whatsapp-lancam-pacote-de-figurinhas-para-as-eleicoes-2022, 2022. Accessed: 2025-04-03.

[144] UNO UNESCO. on genocide prevention, and the responsibility to protect,". *Addressing hate speech on social media: contemporary challenges*, 2021.

[145] Rama Adithya Varanasi, Joyojeet Pal, and Aditya Vashistha. Accost, accede, or amplify: attitudes towards covid-19 misinformation on whatsapp in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[146] Luis Vargas, Patrick Emami, and Patrick Traynor. On the detection of disinformation campaign activity with network analysis. In *Proceedings of the 2020 ACM SIGSAC conference on cloud computing security workshop*, pages 133–146, 2020.

[147] Vítor V Vasconcelos, Sara M Constantino, Astrid Dannenberg, Marcel Lumkowsky, Elke Weber, and Simon Levin. Segregation and clustering of preferences erode

socially beneficial coordination. *Proceedings of the National Academy of Sciences*, 118(50):e2102153118, 2021.

[148] Neeraj Vashistha and Arkaitz Zubiaga. Online multilingual hate speech detection: experimenting with hindi and english social media. *Information*, 12(1):5, 2020.

[149] Otavio R Venâncio, Carlos HG Ferreira, Jussara M Almeida, and Ana Paula C da Silva. Unraveling user coordination on telegram: A comprehensive analysis of political mobilization during the 2022 brazilian presidential election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1545–1556, 2024.

[150] Matteo Vergani, Alfonso Martinez Arranz, Ryan Scrivens, and Liliana Orellana. Hate speech in a telegram conspiracy channel during the first year of the covid-19 pandemic. *Social Media+ Society*, 8(4):20563051221138758, 2022.

[151] Bertie Vidgen and Taha Yasseri. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78, 2020.

[152] Santosh Vijaykumar, Yan Jin, Daniel Rogerson, Xuerong Lu, Swati Sharma, Anna Maughan, Bianca Fadel, Mariella Silva de Oliveira Costa, Claudia Pagliari, and Daniel Morris. How shades of truth and age affect responses to COVID-19 (Mis)information: randomized survey experiment among WhatsApp users in UK and Brazil. *Humanities and Social Sciences Communications*, 8(1), 2021.

[153] Padinjaredath Suresh Vishnuprasad, Gianluca Nogara, Felipe Cardoso, Stefano Cresci, Silvia Giordano, and Luca Luceri. Tracking fringe and coordinated activity on twitter leading up to the us capitol attack. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 1557–1570, 2024.

[154] Shaojung Sharon Wang. More than words? the effect of line character sticker use on intimacy in the mobile communication environment. *Social Science Computer Review*, 34(4):456–478, 2016.

[155] Yuan Wang, Yukun Li, Xinning Gui, Yubo Kou, and Fenglian Liu. Culturally-embedded visual literacy: A study of impression management via emoticon, emoji, sticker, and meme on social media in china. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[156] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, page 19–26, USA, 2012. Association for Computational Linguistics.

[157] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.

[158] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

[159] Derek Weber and Frank Neumann. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining*, 11(1):111, 2021.

[160] Maximilian Wich, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Çakici, and Georg Groh. Introducing an abusive language classification framework for telegram to investigate the german hater community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1133–1144, 2022.

[161] Tomer Wullach, Amir Adler, and Einat Minkov. Towards hate speech detection at large via deep generative modeling. *IEEE Internet Computing*, 25(2):48–57, 2020.

[162] Kai-Cheng Yang, Pik-Mai Hui, and Filippo Menczer. Bot electioneering volume: Visualizing social bot activity during elections. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 214–217, 2019.

[163] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103, 2020.

[164] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First monday*, 2010.

[165] Ping Yi, Ya fei Hou, Yiping Zhong, Shiyong Zhang, and Zhoulin Dai. Flooding attack and defence in ad hoc networks. *JSEE*, 17(2):410–416, 2006.

[166] YouTube. Hate speech policy. https://support.google.com/youtube/answer/2801939, 2025. Accessed: 2025-02-24.

[167] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA, 2018. Association for Computing Machinery.

[168] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*, pages 353–362, 2019.

[169] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer, 2018.

[170] Rui Zhou, Jasmine Hentschel, and Neha Kumar. Goodbye text, hello emoji: Mobile communication on wechat in china. CHI '17, page 748–759, New York, NY, USA, 2017. Association for Computing Machinery.

Appendix A

Dataset Collection Criteria and Keywords

To discover political public groups, we used a set of keywords related to the Brazilian political scenario, initially proposed by [126] and we included updated terms related to relevant people and topics of the period analyzed. These keywords include terms related to relevant political figures during the period analyzed, such as *"Nikolas Ferreira"*, *"Padre Kelmon"*, *"Simone Tebet"*, and *"Alexandre de Moraes"*. Moreover, we added keywords associated with the 2022 presidential election, including: *"Eleições 2022 (election 2022)"*, *"Fraude (fraud)"*, *"Supremo Tribunal Federal (Supreme Federal Court)"*, *"Tribunal Superior Eleitoral (Superior Electoral Court)"*, *"Urna Eletrônica (electronic ballot box)"*, and *"Ditadura Militar (military dictatorship)"*.

Additionally, we included keywords related to the COVID-19 pandemic, such as: *"Pandemia (pandemic)"*, *"Covid-19"*, *"Vacina (vaccine)"*, *"Coronavirus"*, *"Tratamento Precoce (early treatment)"*, and *"Ivermectina (ivermectin)"*.

Appendix B

Rapid Spread Network Threshold Sensitivity

To assess the sensitivity of the 60-second time window threshold in the Rapid Spread Network, we performed an additional analysis to evaluate the impact of varying time windows (30 and 90 seconds). Our goal was to determine whether the chosen time window significantly affects the network structure and to validate the applicability of the methodology to WhatsApp. We applied the methodology summarized in Section 6.1 and Section 6.2 across two different thresholds (30 and 90 seconds).

Creating the rapid spread network with a 60-second time window, we obtained a network with 1,575 nodes, 1,491 edges, and 14,414 coordinated messages (Section 6.3). With a 30-second time window, we have a network with 994 nodes, 918 edges, and 10,465 coordinated messages. For a 90-second time window, the network contains 2,038 nodes, 2,086 edges, and a total of 18,328 coordinated messages.

From these results, modifying the time window from 60 to 90 seconds leads to an increased number of nodes, edges, and coordinated messages. However, the core network structure remains unchanged, with over 70% of the nodes overlapping, indicating that the network is not too sensitive to moderate changes in the time window. Coordinated users identified with a shorter time window remain part of the network structure when a longer threshold is applied. As the threshold rises, the network structure expands, but this does not necessarily lead to identifying more coordinated users. Instead, it may result in more coincidental users sharing the same message due to the larger time interval, highlighting the importance of finding a good balance. Similarly, using a very short time interval may cause many coordinated users to be missed. Based on these findings, the 60-second time window remains a well-supported and balanced choice, suggesting its applicability in the context of WhatsApp.