

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

Gustavo Jota Resende

**Analyzing Information Dissemination in Publicly Accessible Politically  
Oriented WhatsApp Groups**

Belo Horizonte  
2019

Gustavo Jota Resende

**Analyzing Information Dissemination in Publicly Accessible Politically  
Oriented WhatsApp Groups**

**Final Version**

Thesis 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 Master in Computer Science.

Advisor: Jussara Marques de Almeida  
Co-Advisor: Marisa Affonso Vasconcelos

Belo Horizonte  
2019

Resende, Gustavo Jota.

R433a

Analyzing information dissemination in publicly accessible politically oriented whatsapp groups [recurso eletrônico] / Gustavo Jota Resende – 2019.  
1 recurso online (91 f. il, color.): pdf.

Orientadora: Jussara Marques de Almeida.  
Coorientadora: Marisa Affonso Vasconcelos  
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.  
Referências: f. 77-88.

1. Computação – Teses. 2. WhatsApp (Aplicativo de mensagens) – Teses. 3. Desinformação – Teses. 4. Disseminação de informação – Teses. I. Almeida, Jussara Marques de. II. Vasconcelos, Marisa Affonso. III. Universidade Federal de Minas Gerais; Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6\*74(043)



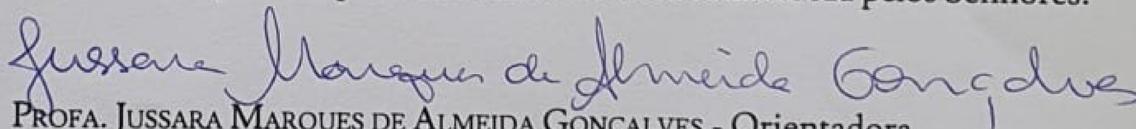
UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS EXATAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

## FOLHA DE APROVAÇÃO

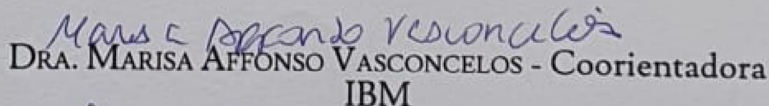
Analyzing Information Dissemination in Publicly Accessible Politically  
Oriented WhatsApp Groups

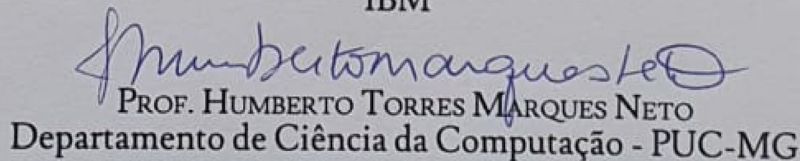
**GUSTAVO JOTA RESENDE**

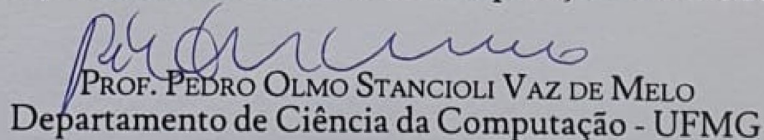
Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:



PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora  
Departamento de Ciência da Computação - UFMG

  
DRA. MARISA AFFONSO VASCONCELOS - Coorientadora  
IBM

  
PROF. HUMBERTO TORRES MARQUES NETO  
Departamento de Ciência da Computação - PUC-MG

  
PROF. PEDRO OLMO STANCIOLI VAZ DE MELO  
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 19 de Dezembro de 2019.

# Acknowledgments

Agradeço primeiramente aos meus pais Enio e Anita, por todo o apoio desde a educação infantil até agora na pós, por sempre me incentivarem a investir na educação e na maior riqueza que é o conhecimento. Aos meus irmãos Livia e Gabriel por sempre me aconselharem de maneira sábia e carinhosa.

Aos meus amigos, sempre presentes nos momentos de lazer, sem nunca deixarem de me apoiar e trazendo momentos leves que foram essenciais nessa jornada. Ao João pelo companheirismo e incentivo que foi essencial em todo o tempo. A Bárbara e o Rafael, amigos e colegas que fiz nesse caminho e tornaram tudo mais divertido. Aos colegas do Locus por todo apoio e trocas, em especial Felipe, Fabiano, Lucas, Átila, Júlio, Carlos e Pablo.

Por fim agradeço as minhas orientadoras Jussara e Marisa por toda paciência e ensinamentos que tornaram possível a conclusão desse projeto. Ao professor Fabrício, e aos colegas Felipe, Júlio e Jonhathan pelas colaborações. Não menos importante, agradeço ao CNPQ pelo apoio financeiro.

# Resumo

O WhatsApp revolucionou a maneira como as pessoas se comunicam e interagem. Não é apenas mais barato que a comunicação tradicional do Short Message Service (SMS), mas também traz uma nova forma de comunicação móvel: as conversas em grupo. Esses grupos são ótimos fóruns para discussões coletivas sobre diversos tópicos. Em particular, em eventos de grande mobilização social, como campanhas eleitorais, as conversas em grupo do WhatsApp são muito atraentes, pois facilitam a troca de informações entre as pessoas interessadas. No entanto, acontecimentos recentes em vários países evidenciaram o uso do WhatsApp para disseminação de mensagens com conteúdo falso e enganoso, levantando preocupações sobre o uso massivo desta plataforma. Motivados por isto, esta dissertação analisa a disseminação de informações no WhatsApp, focando em grupos de temática política acessíveis ao público, coletando as mensagens compartilhadas durante a campanha eleitoral presidencial brasileira de 2018. Nosso estudo contou com um conjunto de dados contendo todas as imagens e mensagens de texto compartilhadas de todos os grupos de temática política durante o período de estudo. Utilizando uma base de dados de informações enganosas previamente verificadas e divulgadas em seis sites brasileiros de checagem de fatos, nós identificamos evidências da presença de mensagens com informação falsa na base de dados analisada. A partir destas evidências, nosso estudo visa identificar características que distinguem as mensagens (de imagem e texto) com conteúdo comprovadamente falso das outras mensagens (com conteúdo não verificado). Para esse fim, analisamos várias propriedades das imagens (por exemplo, conteúdo, principais fontes de imagens e propagação de/para outras plataformas da Web) e do conteúdo textual (por exemplo, uso da linguagem, principais tópicos e sentimento do conteúdo da mensagem), dinâmicas de propagação e estrutura de rede de ambos os conjuntos de mensagens. Identificamos as fontes mais importantes das imagens falsas e descobrimos que, com frequência significativa, elas são postadas primeiro no WhatsApp e depois na Web. Nossas análises também revelaram que mensagens textuais com informações enganosas tendem a se concentrar em alguns tópicos, geralmente carregando palavras relacionadas ao processo cognitivo de *insight*, que caracteriza as mensagens de corrente. Também descobrimos que o processo de propagação é muito mais viral para imagens e mensagens de texto com conteúdo falso.

**Palavras-chave:** Grupos de WhatsApp, Desinformação, Disseminação de informação, Imagens, Informação textual.

# Abstract

WhatsApp has revolutionized the way people communicate and interact. It is not only cheaper than the traditional Short Message Service (SMS) communication but it also brings a new form of mobile communication: the group chats. Such groups are great forums for collective discussions on a variety of topics. In particular, in events of great social mobilization, such as electoral campaigns, WhatsApp group chats are very attractive as they facilitate information exchange among interested people. Yet, recent events in several countries have highlighted the use of WhatsApp to spread messages with false and misleading content, raising concerns about the massive use of this platform. This master thesis analyzes information dissemination within WhatsApp, focusing on publicly accessible political-oriented groups, collecting all shared messages during the 2018 Brazilian presidential campaign. Our study relied on a dataset containing images and textual messages shared in political groups during the study period. We identified the presence of misinformation in the contents of these messages using a dataset of priorly checked misinformation from six Brazilian fact-checking sites. From this evidence, our study aims to identify characteristics that distinguish messages (image and textual) with proven false content from other messages (with unverified content). To that end, we analyzed various properties of the images (e.g., main content, main image sources, and propagation from/to other Web platforms) and of the textual content (e.g., language usage, main topics and sentiment of message's content), propagation dynamics and network structure of both message sets. We identify the most important sources of the fake images and found that they much more often appear first on WhatsApp and then on the Web. Our analyses also revealed that textual messages with misinformation tend to be concentrated on fewer topics, often carrying words related to the cognitive process of *it insight*, which characterizes chain messages. We also found that their propagation process is much more viral for both images and textual messages.

**Keywords:** WhatsApp groups, Misinformation, Information dissemination, Images, Textual information.

# List of Figures

3.1	WhatsApp Data Collection Flowchart . . . . .	30
3.2	Images Filtering Flowchart . . . . .	32
3.3	Images Filtering Flowchart . . . . .	33
3.4	Numbers of daily messages with media content shared on all groups . . . . .	34
4.1	Images checked as fake by both fact-checking methodologies. . . . .	38
4.2	Flowchart of an Automatic Methodology for Finding Misinformation in Images	39
4.3	Distributions of image categories. . . . .	42
4.4	Most popular domains for images shared on WhatsApp publicly accessible groups. . . . .	43
4.5	Most popular Twitter accounts for images shared on WhatsApp publicly accessible groups. . . . .	44
4.6	Flowchart of Methodology for Finding Misinformation in Textual Messages . .	44
4.7	Distributions of message sizes . . . . .	47
4.8	LIWC attributes that occur more frequently in messages with misinformation.	48
4.9	Sentiment polarity of messages. . . . .	50
4.10	Flowchart of Methodology for characterizing WhatsApp textual messages in topics . . . . .	51
4.11	Distributions of topics inferred by LDA. . . . .	53
4.12	Word clouds of the top 500 words (translated to English). . . . .	54
4.13	Word tree for the word root <i>Please</i> . . . . .	55
4.14	Most frequent domains in textual messages. . . . .	56
5.1	Cumulative distributions of the reach of each image in terms of distinct users, distinct groups and total shares. . . . .	59
5.2	Cumulative distributions of the reach of each textual message in terms of distinct users, distinct groups and total shares. . . . .	60
5.3	Distributions of lifetimes of messages with misinformation and with unchecked content . . . . .	62
5.4	Distribution of burst times for messages with misinformation and messages with unchecked content . . . . .	64
5.5	Distribution of inter and intra-group for images with misinformation and messages with unchecked content. . . . .	65



## List of Figures

---

5.6	Distribution of inter and intra-group for textual messages with misinformation and messages with unchecked content. . . . .	66
5.7	Cumulative Distribution Function for Temporal Propagation on Web of WhatsApp Images . . . . .	67
5.8	Network representation of images shared on WhatsApp and on the Web. . . .	68
5.9	WhatsApp Network of Users with public political groups in common . . . . .	70
5.10	WhatsApp Network of Users with public political groups in common . . . . .	71
5.11	The Misinformation Network of WhatsApp Images . . . . .	73

# List of Tables

3.1	Overview of our dataset. . . . .	31
4.1	Sharing of images on monitored WhatsApp groups. . . . .	37
4.2	Overview of images shared during election campaign period: misinformation versus unchecked content. . . . .	40
4.3	Image Categories . . . . .	41
4.4	Image Categories <i>Fleiss's</i> $\kappa$ . . . . .	41
4.5	Topics inferred by LDA algorithm. . . . .	51
5.1	Network metrics for WhatsApp graphs. . . . .	70
A.1	Dictionary of words related to the 2018 Brazilian elections - Part 1 . . . . .	90
A.2	Dictionary of words related to the 2018 Brazilian elections - Part 2 . . . . .	91
A.3	Dictionary of words related to Ideologies and Political Sides . . . . .	92

# Contents

<b>1</b>	<b>Introduction</b>	<b>12</b>
1.1	Motivation . . . . .	12
1.2	Goals . . . . .	14
1.3	Contributions . . . . .	15
1.4	Outline . . . . .	16
<b>2</b>	<b>Background</b>	<b>18</b>
2.1	Main Concepts . . . . .	18
2.1.1	Online Information . . . . .	18
2.1.2	WhatsApp Platform . . . . .	19
2.2	Related Work . . . . .	20
2.2.1	Online Information Diffusion . . . . .	21
2.2.2	Misinformation Properties . . . . .	22
2.2.2.1	Static Properties . . . . .	22
2.2.2.2	Dynamic Properties . . . . .	24
2.3	Studies on WhatsApp . . . . .	26
<b>3</b>	<b>WhatsApp Data Collection</b>	<b>29</b>
3.1	Methodology . . . . .	29
3.2	Dataset . . . . .	30
3.3	Data Filtering . . . . .	31
3.3.1	Images . . . . .	32
3.3.2	Textual Messages . . . . .	33
3.4	Overview . . . . .	33
<b>4</b>	<b>Image and Textual Content Properties</b>	<b>36</b>
4.1	Images . . . . .	36
4.1.1	Misinformation . . . . .	37
4.1.1.1	Labeling with a Fact-Checking Agency . . . . .	37
4.1.1.2	An Automatic Methodology for Finding Misinformation . . . . .	38
4.1.2	Content Labeling . . . . .	40
4.1.3	WhatsApp Images in other Websites . . . . .	42
4.2	Textual Messages . . . . .	43
4.2.1	Identifying Misinformation . . . . .	44

4.2.2	Textual Properties . . . . .	46
4.2.2.1	Message Sizes . . . . .	46
4.2.2.2	Psychological Linguistic Features . . . . .	47
4.2.2.3	Sentiment Analysis . . . . .	49
4.2.2.4	Topic Analysis . . . . .	50
4.2.2.5	Frequent Terms . . . . .	52
4.2.2.6	Domains Shared . . . . .	55
4.3	Summary of Results . . . . .	56
<b>5</b>	<b>Propagation Dynamics</b>	<b>58</b>
5.1	Message Reach . . . . .	58
5.2	Propagation Within WhatsApp . . . . .	61
5.2.1	Lifetimes . . . . .	62
5.2.2	Burst Times . . . . .	62
5.2.3	Intra and Inter Group Times . . . . .	63
5.3	Propagation to and from the Web . . . . .	66
5.4	Network Structures . . . . .	69
5.4.1	General Network Properties . . . . .	69
5.4.2	Misinformation Network . . . . .	72
5.5	Summary of Results . . . . .	72
<b>6</b>	<b>Conclusions and Future Work</b>	<b>75</b>
	<b>Bibliography</b>	<b>78</b>
	<b>Appendix A Dictionary of words related to the 2018 Brazilian elections</b>	<b>90</b>

# Chapter 1

## Introduction

WhatsApp is a world-wide popular messaging app with more than 1.5 billion active users which is currently the main messaging app in many countries, including India, Brazil, and Germany [13]. Nearly everyone with a smartphone uses WhatsApp in Brazil (about 120 million active users) to keep in touch with friends and family, do business, as well as read the news [57].

WhatsApp changed how people communicate when using smartphones, with a simple and easy-to-use interface, the app allows its users to exchange textual and multimedia messages in private and group conversations. There are key features in WhatsApp that make this app unique. First, any communication within the app is end-to-end encrypted, meaning that messages, photos, videos, voice messages, documents, status updates, and calls are only seen by those involved in the communication. Second, WhatsApp allows users to easily create and organize chat groups. The conversation in groups allows users to chat and interact instantly with all of those who joined the group. These groups, which are limited to 256 members, are by default private, as group administrators decide who can join them. However, a group manager may choose to share the link to join it on websites or social networks. In such a case, anyone with access to the link can join the group, which becomes, from a practical perspective, publicly accessible. Finally, WhatsApp provides features for viral spreading, allowing users to broadcast an initial message to 256 contacts or groups or forward content to 20 contacts or groups<sup>1</sup>.

### 1.1 Motivation

While the emergence of online platforms for communication and information dissemination has generated a more connected world, it has also had serious negative implications that need to be analyzed and addressed, such as the dissemination of false

---

<sup>1</sup>The message forwarding was limited to up to 5 groups in India and 20 in the rest of the world along the period this work was developed. Currently, the limit has been updated to 5 worldwide.

information [53]. WhatsApp groups facilitate the dissemination of different types of content including chain messages, news, memes, and rumors, including the so-called fake news. These fabricated stories end up creating confusion in the public about the truthfulness of current events, posing a troubling challenge for the news industry as well as for society as a whole.

In fact, WhatsApp groups have been used out as one of the main tools for content dissemination in the recent past, as noted in May 2018 during the Brazilian trucker strike, when the app facilitated the mobilization of thousands of them [83]. Moreover, recent events have raised serious concerns that WhatsApp can become a fertile ground for groups interested in disseminating misinformation, especially as part of articulated political campaigns. In addition, in India and Brazil where the app already reached 200 and 120 million users respectively, as reported [20, 54], the spread of misinformation in WhatsApp has had consequences for society in those both countries. In 2018, unfounded allegations disseminated over WhatsApp have fueled mob lynchings in India that killed more than 20 people in a two-month window [32]. The 2018 Brazilian elections experienced an information war organized within WhatsApp where false rumors, manipulated photos, decontextualized videos, and audio hoaxes have become campaign ammunition and went viral on the platform with no way to monitor their full reach or origin [57].

Indeed, WhatsApp has acknowledged the importance of reducing the spread of misinformation by restricting the number of times a unique message can be forwarded by the same user<sup>2</sup>. This is the first step to constrain the spread of fake news. Yet, given the great popularity of the application, its effectiveness is naturally limited. A recent study showed how those current efforts deployed by WhatsApp are ineffective in blocking the propagation of misinformation campaigns in public groups [62]. We here are particularly interested in features of *misinformation*, which refers to reportedly false (or inaccurate) information. It is of utmost importance to identify characteristics of messages containing misinformation that distinguish them from regular content, as a step to build effective countermeasures against their dissemination.

Previous studies about WhatsApp focused on understanding the general patterns of how users interact with the application [30, 9, 8] as well as its use on specific tasks (e.g., educational tasks, medical information exchange) [105, 7] and misinformation spread [11, 74, 18]. Yet, no prior study, has studied the dissemination of *images* in publicly accessible politically oriented groups in WhatsApp, highlighting some differences in images containing previously identified misinformation from the rest. Also no previous work focused on exploring the presence of misinformation in *textual* messages, which are the most common types of content shared in the system, and whether there are particular content features that distinguish them from the other textual messages.

---

<sup>2</sup><https://www.theguardian.com/technology/2019/jan/21/whatsapp-limits-message-forwarding-fight-fake-news>

Thus, the dissemination of information on social media is particularly potent and dangerous for two reasons: the propagation speed and the viral behavior. While extremely important, detecting misinformation is a technically challenging problem. Many techniques for the automatic detection of false news have been proposed in the literature [90, 104]. Although each work provides a different set of attributes to be observed, the main ones to be addressed are related to (1) the text of an article, (2) the responses of the users who receive it and (3) the users who disseminate it.

We here give a first step on understanding information dissemination on WhatsApp publicly accessible politically oriented groups, where we investigate the most frequent media types on messages, the main topics, content and how information interplay between WhatsApp groups and other Web platforms. By doing this, we quantify the presence of misinformation in the content of those groups and present a characterization of the differences between misinformation and other messages whose content was unchecked for image and textual messages.

## 1.2 Goals

Based on these characteristics, a number of related issues arise to identify how information is disseminated in public WhatsApp groups. In this master thesis we intend to provide a large scale investigation of information dissemination within *WhatsApp public groups*. We focus on political-oriented public accessible groups as we expect greater user engagement in topics of stronger social impact. By doing so, we offer a first look into *misinformation* dissemination within *WhatsApp*. We also aim to compare the images and textual messages containing previously reported misinformation with other messages whose content was unchecked. We characterize these two sets of messages in terms of language usage, the main topics and sentiment of the message's content, as well as their propagation dynamics. More specifically, we here tackle the following research questions.

**RQ1:** What kind of content is shared in public groups in WhatsApp? Is there fake news in these messages?

**RQ2:** What are the differences in terms of features between textual messages and images containing misinformation and the rest?

**RQ3:** How are the propagation dynamics of the messages containing misinformation (i.e., how long they remain being spread, how many people and groups spread them) and how it differs from the propagation of other messages?

To answer these questions, we first identify publicly accessible groups related to Brazilian politics in WhatsApp, by searching the Web and other social networks such as

Twitter and Facebook for invitation links to WhatsApp groups. These groups are suitable for activism and political engagement, making them a potential target of misinformation campaigns that might attempt to maximize the audience of a story with misinformation by sharing it with people that are engaged in supporting political candidates. We joined those groups and gathered the content shared within them for a time period corresponding to the first round of the 2018 Brazilian general elections campaign (August 16<sup>th</sup> to October 7<sup>th</sup>, 2018), with 364 groups monitored.

## 1.3 Contributions

The main contributions of this master thesis are:

- **Overview of WhatsApp Information Dissemination**

We investigate and analyze the content shared and the user interactions within the monitored WhatsApp groups to understand how users disseminate information in such environments. Our results show that images are often the most shared type of media, and they usually carry satires, news, and activism-related content. We also show that WhatsApp has a network configuration similar to many other online social networks (e.g., Twitter or Facebook) which connects thousands of users. Thus, it has the potential to make any information become viral. Moreover, we analyze how the content dissemination crosses the boundary between WhatsApp and other Web platforms.

- **Characterization of Misinformation on Images**

We explore the presence of misinformation campaigns in the monitored groups. We identify misinformation in image content by relying on two sources: (i) a Brazilian fact-checking agency; and (ii) a proposed automatic procedure that exploits the results of Google searches to identify images that appear in well-known fact-checking websites. Our results show a considerable number of images checked as containing misinformation, which was largely disseminated in the monitored groups. We assess the occurrence of such images in Web domains and Twitter accounts.

- **Characterization of Misinformation on Textual Messages**

In addition, we also analyze misinformation on textual messages, where we gathered fake news from six Brazilian fact-checking agencies and used it to identify misinformation in the textual messages of our collected dataset. Our analyses unveiled a number of interesting findings regarding the dissemination of misinformation in



textual messages in the WhatsApp groups monitored. We found that messages with misinformation tend to be slightly smaller (especially in the number of words), partially due to the larger presence of URLs in their contents. Moreover, they tend to be concentrated on fewer topics, often carrying words related to the cognitive process of insight (which characterizes chain messages).

- **Misinformation Propagation Dynamics**

Finally, we also analyze the propagation dynamics of textual messages and images. We contrasted images and textual messages with misinformation and found that their propagation process is much more viral, reaching a larger number of users and groups. However, with a distinct behavior: images with misinformation have a lower time interval between consecutive shares but are longer for misinformation in textual messages. Also, textual messages with misinformation tend to propagate faster within particular groups, whereas images tend to take faster to propagate across different groups. Moreover, by comparing the timestamps when an image first appeared on WhatsApp (as captured by our data) and on other Web applications, we find that WhatsApp was the primary source of 30% of the identified images containing misinformation. Further, we study the network that emerges from the sharing of images. Our analyses reveal that a few groups are the most responsible for disseminating images with misinformation.

The results of this master thesis are summarized in two papers [75, 74]. Moreover, we started our study in WhatsApp groups in [76] and we also participated in the design of the WhatsApp Monitor<sup>3</sup> [61], a Web-based system to help the top Brazilian official fact-checking agencies and journalists. The tool displays the most popular content shared in the monitored publicly accessible groups on a daily basis.

## 1.4 Outline

The remainder of this work is organized as follows. Chapter 2 reviews background information and existing literature on misinformation and studies on WhatsApp. Chapter 3 describes the data collection process and the creation of our WhatsApp dataset used in our analysis. In Chapter 4, we present our analyses where we characterize the content properties of WhatsApp on images and textual messages. The propagation dynamics of image and textual messages within each group as well as across different groups on

---

<sup>3</sup>WhatsApp Monitor: <http://www.monitor-de-whatsapp.dcc.ufmg.br/>

WhatsApp also follows in Chapter 5. Finally, conclusions and directions for future work are presented in Chapter 6.

# Chapter 2

## Background

Many prior studies were carried out with the purpose of analyzing information (in particular fake news) dissemination in social networks, studying techniques to identify misleading information and entities that are active in the creation and dissemination of content. Thus, in this chapter, the main concepts and prior investigations used as a basis for our study are presented. The chapter is divided in two main parts. Section 2.1 presents the main concepts related to the dissemination of information (notably false information) online and the WhatsApp platform. Section 2.2 discusses prior studies on information dissemination in general and false news propagation in particular. Finally, we also highlight some previous work that focused on WhatsApp in Section 2.3.

### 2.1 Main Concepts

#### 2.1.1 Online Information

The Web provides an environment where information can be spread in a small period of time reaching millions of readers and viewers. Moreover, new social technologies facilitate rapid information sharing and large-scale information cascades [103]. However, the easy access to information on the Web has led to increased visibility and impact of both true and false information. In particular, false information on the web and social media has affected political candidates campaign [57], public health [107].

False information has been further specified based on the intent (or not) of its author as misinformation and disinformation [39]. The term *misinformation* is often used to refer to false information built with no purpose to deceive. Frequent causes of misinformation are contortion of facts or accurate information led by cognitive biases or absence of perception [49]. An example of a misinformation is the sharing of news about

an accident that happened, but with wrong information about the victims, caused by a misunderstanding, of what actually happened.

Disinformation is defined as a piece of false information built with the purpose to deceive people [68]. Most of the disinformation campaigns focus on influencing public opinion or reaching website hits, and thus become profitable from web advertising. Recent examples are political disinformation spread during 2016 USA presidential elections [35] and during 2018 Brazilian presidential elections [57]. During election campaigns, people may produce false content, with the intent to deceive, to support a particular candidate and/or also harm the opposition.

In this master thesis, we do not analyze the intention of those who created a particular piece of false information. Therefore, we adopt the most general term, *misinformation*, to refer to all pieces of information that has been reported as false by official fact-checking agencies.

Information can also be characterized based on knowledge, as either opinion-based or fact-based [99]. When information is based on individual opinion and there is no universal truth, it is opinion-based information. The writer's opinion may create a false belief to affect or influence other people's decision. An example of false information that is opinion-based, is when a fake profile of a celebrity on a social platform post some fake opinion about some fact or event.

A piece of fact-based information may be false even when the information characterizes a true fact but it is not entirely accurate. The motive of this type of information is to make it harder for the reader to distinguish true from false information, and make them believe in the false version of the information [68]. This type of false information includes fake news, rumors, and fabricated hoaxes. Facts may be checked and confirmed (or not). Thus, a false fact-based information may be debunked and proven false. An example of false information that is fact-based, is a real event described with false details of how it occurred. In this master thesis, we focus on fact-based misinformation, by relying on third-party fact-checking to categorize it.

## 2.1.2 WhatsApp Platform

WhatsApp Messenger is a cross-platform messaging application released in 2009, which has become popular mainly in recent years. Initially, WhatsApp was an instant-only communication tool, for mobile devices, but now it can be used on both mobile devices (smartphones and tablets) and on personal computers through a Web version, WhatsApp Web. The application provides a variety of communication features like text

messaging, exchange of photos, audios, videos and files in general as well as the option to make phone and video calls.

WhatsApp messages, voice and video calls between a sender and receiver that use WhatsApp client software released after March 31, 2016 are end-to-end encrypted<sup>1</sup>. This end-to-end encryption protocol was designed to prevent third parties and WhatsApp from having plaintext access to messages or calls.

The chat feature on WhatsApp allows users to have private and group conversations. In the first, the user can exchange messages with a contact, added by their mobile number, in a private conversation. The group chats, have a name and a description and allows up to 256 people. The group admin can manually add members, or generate a link to invite people to join the group. Those invite links of WhatsApp groups may be posted and shared on social networks and well-known websites and are typically themed around particular topics, like politics, sports, tv shows and music. As invite link to a WhatsApp group is shared online, anyone that finds it can easily join the group by simply clicking on it. Thus, the ability of a group admin to share invite links effectively makes the group publicly accessible.

One of the great financial advantages of WhatsApp is its cost: message exchange and calls are free, requiring only an internet connection. Thus the user can make use of an existing internet connection in the environment where he is, such as schools and public places. More recently, WhatsApp has also become a powerful tool to influence people during political campaigns, especially in countries in South America, Africa, and Southeast Asia [15, 60]. This was observed in Brazil, where family groups were responsible for 51% of the dissemination of fake news on WhatsApp during the period of the 2018 presidential elections [34]. Motivated by this, in this master thesis, we focus on studying publicly accessible politically oriented WhatsApp groups.

## 2.2 Related Work

In this section we present studies that are related to the main topics of this master thesis. We start by presenting in Section 2.2.1 a brief review of important studies on information dissemination in online social networks. Section 2.2.2 focuses on misinformation and fake news. Finally, in Section 2.3 we review some recent studies on WhatsApp.

---

<sup>1</sup>WhatsApp Security available on: <https://www.whatsapp.com/security/>

### 2.2.1 Online Information Diffusion

Online social networks play a major role in the diffusion of information by increasing the spread of novel information and diverse viewpoints [4]. Many studies focused on understanding how information spreads on different social networks by exploring two main categories of models. On one hand, some prior studies focused on developing explanatory models [80, 17, 85], which aim to infer the underlying spreading cascade, given a complete activation sequence. Others have investigated predictive models [29, 36], that aim at predicting how a specific diffusion process will unfold in a given network, from both a temporal and spatial perspectives, learning from past diffusion traces [37].

Gomez *et al* proposed a time-varying inference algorithm, *INFOPATH*, that uses stochastic gradients to provide on-line estimates of the structure and temporal dynamics of a network that changes over time [80]. Based on experimentations on Twitter data, Choudhury *et al* concluded that sampling methods that consider both network topology and users' attributes such as activity and localisation allow to capture information diffusion with lower error in comparison to naive strategies, like random or activity-only based sampling [17]. Sadikov *et al* developed a method based on a ktree model designed to, given only a fraction of the complete activation sequence, estimate the properties of the complete spreading cascade, such as its size or depth [85]. Galuba *et al* proposed a model that relies on parameters such as information virality, pairwise users degree of influence and user probability of adopting any information [29]. Guille *et al* modeled the propagation process as asynchronous independent cascades, with parameters that are estimated from social, semantic and temporal nodes' features using logistic regression [36].

Yang *et al* proposed a social role-aware information diffusion, the *RAIN* model, which integrates social role extraction and diffusion modeling into a unified framework [110]. Hoang *et al* predicted whether a post is going to be forwarded or not and how much it is going to be diffused. The authors concluded that the number of followers, and the number of groups that the user belongs to are the most important features for prediction effectiveness [40]. Jung *et al* detected information diffusion across the boundaries of Twitter and news sites to understand how media outlets are perceived on social media, and found that information diffusion occurs continuously across platform boundaries. They also reported that social media communication is more likely to be picked up by the news coverage and thus has a greater potential effect on the media agenda than the news sites do on Twitter [44].

## 2.2.2 Misinformation Properties

### 2.2.2.1 Static Properties

Fake news has encountered a suitable means for fast, cheap, and easy dissemination in social media systems. Indeed, these platforms have been the main vehicle for public opinion manipulation and fake news dissemination [103, 77, 73]. The misinformation spreading process starts with small communities of individuals who engage with questionable publishers and share them as rumors [84]. A recent investigation on Facebook [116] exposed the formation of a phenomenon named *echo chambers* where users interact with like-minded people sharing the same system of beliefs. The authors found that the size of misinformation cascades may be approximated by the same size of the echo chamber. Several other recent studies have investigated how online social networks may impact many global political scenarios, such as the White Helmets in the Syria [93] and the 2016 US presidential campaign [27, 14, 89]. Studies during the 2016 U.S. presidential election campaign observed a strong correlation of the number of visits to fake news websites (i.e., sites that deliberately publish hoaxes and misinformation) and aggregate voting patterns at state and county levels [27]. Cunha *et al.* studied the perception and the conceptualization of the term *fake news* in the traditional media using eight years of data collected from news outlets based in 20 countries [14]. They found that the interest for the term *fake news* suddenly increased after the 2016 US election and this growth was accompanied by a change of framing around the term *fake news* from, for instance, topics regarding the media industry itself to those related to political affairs. Shao *et al.* presented an in-depth analysis of the misinformation diffusion network on Twitter in the run-up to and wake of the 2016 US presidential election [89]. The authors found that the network is strongly segregated along with the two types of information circulating in it and that dense, stable core emerged after the elections. They also characterized the main core in terms of multiple centrality measures and proposed efficient strategies to reduce the circulation of information by penalizing key nodes in this network.

Social bots are one of the most common types of manipulation attacks, emulating real users, posting content and interacting with real users and other bots [25, 63, 5]. They have been used as political advocates on Twitter [79] during debates. Another misinformation campaign was observed during the 2017 French presidential election, in which bots' posts with unauthentic documents about a candidate quickly spread on Twitter two days before the final voting [24]. Facebook was also a target of misinformation spread aiming at influencing American voters during the 2016 presidential campaign. Using Facebook

Ads platform, groups linked to the Russian Intelligence Research Agency (IRA) bought about 3,000 ads linked to 470 user accounts targeting voters from the swing states [77, 47]. Since then, Facebook has performed measures to mitigate fake news dissemination such as removing fake accounts related to political movements and working directly with fact-checking websites [55]. However, fake news is not disseminated exclusively by social bots or through ads but also by real users. A study analyzed over 126,000 cascades of fact-checked news stories on Twitter and found false news was 70% more likely to be retweeted than the true stories. The most surprising was that humans are more likely to spread false news than bots [103]. Studies on Twitter showed that polarization increases when content (e.g., URL, hashtags) is related to fake news and users tag news and statements that they disagree or that are considered by the opposition groups as fake [78].

Some authors have proposed learning methods to automatically detect fake messages ranging from lexical to deep learning approaches exploring linguistic and network features [90, 104]. However, Zhang *et al.* reported that the detection of misleadingness content on misinformation requires an assessment of the content, context, literal meaning and intentions in order to determine the utterer's meaning, hence the implicature (if any is present) and to work it out [113]. S oe *et al.* reported that producing ground truth is the critical issue when developing machine learning models for predicting misinformation [91]. In this kind of classification task, textual features are great semantic resources used very often on many approaches that explore the language structure [12, 70, 33], sentiment and other psycho-linguistic cues [70, 103], topic models [43, 42] and even political biases [3] of the messages.

Other studies focused on fighting the spreading of misinformation. Zhang *et al.* proposed a model to fight the misinformation spread solving the *Distance-constrained Misinformation Combat under Uncertainty* problem, which aims to both reduce the spread of misinformation and enhance the spread of correct information within a given propagation distance [114]. The misinformation containment problem aims at limiting the spread of misinformation in online social networks by launching competing campaigns and was studied in [100], where the authors provided a formal model to address the MC (Misinformation Combat) problem from the view of combinatorial optimization. The authors showed that the MC problem can be close to submodular optimization problems and designed an evaluated an effective algorithm for solving it. An efficient online algorithm, named *Curb*, that leverages the crowd to detect and prevent the spread of misinformation in online social networking sites was presented in [46]. The algorithm selects stories to send for fact-checking and when to do. The authors evaluated the algorithm using two real-world datasets gathered from Twitter and Weibo and showed that the algorithm may be able to effectively reduce the spread of misinformation.

As polarization plays an important role in the misinformation spreading process, Del Vicario *et al.* presented a general framework for timely identification of polarizing



content that enables to predict the topics of future fake news on social media, and built a classifier for fake news detection [102]. The authors validated the performance of their method on a massive dataset of official news and hoaxes on Facebook. Popat *et al.* created *CredEye*, a system for automatic credibility assessment that takes a natural language claim as input from the user and automatically analyzes its credibility by considering relevant articles from the Web [69].

All these prior efforts, however, focused mostly on news articles and posts in online social networks such as Facebook [95, 73], Twitter [42, 12] and Weibo [104]. We are aware of only one work that analyzed the dissemination of misinformation on WhatsApp, which was developed concurrently with this master thesis. We defer a discussion of this work to Section 2.3.

WhatsApp owns peculiarities that differ it from other platforms. For instance, WhatsApp groups are fundamentally chat rooms where any member can share a piece of content instantly reaching all other members. Unlike other social networks, WhatsApp groups form somewhat small communities<sup>2</sup> where content dissemination is driven solely by the members' intentions, with no influence of any recommendation or news feed algorithm. Thus, information spread in such environment may convey particular properties worth studying.

### 2.2.2.2 Dynamic Properties

Misinformation has a viral behavior, depending on how the false information spreads, inspiring many authors to study and analyze their propagation properties. Kumar *et al.* studied hoax articles, *i.e.*, articles containing fabricated facts about nonexistent entities or events on *Wikipedia* by measuring how long they survive before being debunked, how many page views they receive, and how heavily they are referred to by documents on the Web [50]. The authors found that most hoaxes are detected quickly and have little impact on Wikipedia, while a small number of hoaxes survive long and are well-cited across the Web.

When studying the propagation of misinformation authors have found that usually a few users are behind the advance of false information. Gupta *et al.* showed the top thirty users out of 10,215 users (0.3%) resulted in 90% of the retweets of fake images during the Hurricane Sandy (2012) [38]. Shao *et al.* also reported that fake news are dominated by a few very active users, while fact-checking is a more grass-roots activity, done several hours later [88]. However, Zubuiaga *et al.* argued that the prevalent tendency for users

---

<sup>2</sup>The number of members in a group is limited to 256.

is to support every unverified rumor [117].

Several studies also reported that false information spreads deeper compared to real information. Friggeri *et al.* found that rumor cascades run deeper on *Facebook* than reshare cascades of real news in general, being more viral. This observation is consistent with the propagation relying less on who posts the rumor, and more on the highly contagious nature of the rumor itself [28]. Also, Zeng *et al.* contrasted rumor-affirming messages with rumor-correcting messages on Twitter and showed that information related to rumors, both supportive and denying them, spread faster than non-rumors [112]. In [19], the authors conducted simulations to show that even a rumor that started at a random node of the Twitter network on average reached 45.6 million of the total of 51.2 million members within only eight rounds of communication. Researchers also reported that misinformation propagates faster during the initial phases.

Zubiaga *et al.* explored the alternative media ecosystem through a Twitter lens and found that the spread of false information occurs largely before it is even debunked [117]. Vosoughi *et al.*, in turn, investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017 and reported that falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information [103]. In [12], the authors analyzed the credibility of information shared on Twitter, discovering that there are measurable differences in the way credible and not credible messages propagate, whereas the authors of [103] showed that fake news tends to spread faster than the real news on the platform.

In this master dissertation, we also contrast the propagation dynamics of messages containing previously checked misinformation with the rest within WhatsApp publicly accessible politically oriented groups. We also analyze characteristics of the propagation dynamics of these messages analyzing the total time a message remains being shared in the platform, and the time between consecutive shares of the same content.

As reported, many researchers studied misinformation propagation focusing on one particular social network or application. In contrast, a few others have results on how misinformation spreads across many social applications. For example, Zannettou *et al.* filled this gap by studying mainstream and alternative news shared on Twitter, Reddit, and 4chan and reported that alt-right communities within 4chan and Reddit can have a surprising level of influence on Twitter, providing evidence that communities often succeed in spreading alternative news to mainstream social networks and the greater Web [111]. They also reported that users on different platforms prefer different news sources, especially when it comes to alternative ones. The author of [2] found that that only 60% of incoming traffic from a sample of leading fake and hyper-biased news sites seemed to be coming out of Facebook and Twitter and the remaining 40% of web traffic was

organic, coming from direct website visits, peer-to-peer shares, text/instant messaging, e-newsletters subscriptions, RSS, and search engines. In this master thesis, we also investigate how information from WhatsApp publicly accessible politically oriented groups crosses the boundaries to and from the Web analyzing the domains and Twitter profiles where the information also appeared as well as the time intervals between the first appearances of a particular piece of information on WhatsApp (on the monitored groups) and elsewhere on the Web.

Other authors have analyzed the network characteristics of rumor spreading. For example, Subrahmanian *et al.* analyzed the network of users on Twitter, observing that some bot accounts that spread false information are close to each other and appear as groups in Twitter’s follower-followee network, with significant overlap between their followers and followees [94]. Bessi and Ferrara [5], in turn, inferred political partisanship from hashtag adoption, for both humans and bots, and studied spatio-temporal communication, political support dynamics, and influence mechanisms by discovering the level of network embeddedness of the bots. The authors observed that bots become increasingly central in the rebroadcasting network. Those results showed that the presence of social media bots can indeed negatively affect democratic political discussion rather than improving it, which in turn can potentially alter public opinion and endanger the integrity of a democratic action (e.g., election campaign). Starbird [92] used tweeted URLs to generate a domain network, connecting domains shared by the same user, then conducted a qualitative analysis to understand the nature of different domains and how they connect. The author found that those domains form tightly connected clusters, meaning that many users mention these domains together in their false information tweets.

In this master thesis, we analyze the network structure of WhatsApp publicly accessible politically oriented groups. We show how some groups are interconnected by common members and that, despite not being designed to be a social network, WhatsApp does have network properties common to other online social networks that favor content virality. We also create a network of misinformation, presenting the groups where misinformation is first sent, and the groups where it appears more frequently.

## 2.3 Studies on WhatsApp

WhatsApp’s recent worldwide popularity has motivated authors to study the platform. Fernandez-Robin *et al.* studied the intentions that drive users when using WhatsApp. By applying a questionnaire to 579 individuals, they observed that people in general use WhatsApp for leisure and entertainment, although others also manifested using it for

work, study, and informative reasons [22]. There are also others reports that WhatsApp is being massively used not only as an important tool for marketing<sup>3</sup> but also as a vehicle for spreading fake news [14, 10]. A large number of users subscribe to WhatsApp groups that are in alignment with their ideology, thus receiving content that usually reinforces their biases [21]. For example, Cunha *et al.* pointed WhatsApp as one of the leading sources of misinformation spreading, showing how users are easily manipulated through the spread of misleading information [14]. In a study with two large datasets of news, Caetano *et al.* analyzed the public perception of the content often associated with WhatsApp in different regions of the world and over time [10]. They concluded that the vocabulary and topics around the term *WhatsApp* in the media have been changing over the years and recently concentrate on matters related to misinformation, politics, and criminal scam.

Some studies [105, 7] focused on understanding how users interact using Whatsapp for performing different tasks such as educational tasks, medical information exchange, etc. Gazit and Aharony investigated the prediction of the level of participation of 130 students in Israel in WhatsApp groups by performing a survey with the students [31]. Their findings confirmed that psychological factors such as social support, extroversion, and narcissism are significant factors for the prediction of the level of participation in WhatsApp groups. Marfianto and Riadi, in turn, used Digital evidence, obtained using forensic analysis procedures and used textual mining techniques to identify messages from crime perpetrators on WhatsApp [58]. Al Khaja *et al.* analyzed a WhatsApp dataset, collected from a personal smartphone, and concluded that the majority of the drug-related messages were potentially misleading or false claims that lacked credible scientific evidence [1].

Some other studies have investigated how users behave as they share messages in WhatsApp, particularly within chat groups. Rosenfeld *et al.* found that the younger users use WhatsApp more frequently, and women use this network more often than men to communicate in general and with relatives while men, on the other hand, are generally members of larger communication groups and send shorter messages [82]. Schwind and Seufert developed *WhatsAnalyzer*, a web-based tool to collect and analyze chat histories of the mobile messaging application WhatsApp, where some visual data is produced to show basic insights into their communication [86]. As a side effort in the development of this master thesis, we also participated in the design of the WhatsApp Monitor [61], a web-based system that helps researchers and journalists explore the nature of content shared on WhatsApp public groups from three different contexts: Brazil, India and Indonesia. Garimella and Tyson proposed a generalizable data collection methodology for WhatsApp public groups [30], whereas Seufert *et al.* investigated the emerging group-based communication paradigm on WhatsApp and its implications on mobile network traffic [87]. Caetano *et al.* developed a script to collect all messages from WhatsApp Web

---

<sup>3</sup><http://nyti.ms/2L3AV3M>

and analyzed user behavior in public groups using a three-layer hierarchical approach (e.g., message, user and groups) [9]. An epidemiological SIR model to describe dynamics of spreading of fake news through WhatsApp was presented in [45], where they reported that the rate of growth of misinformation through the social media platform is increasing every year by the age group of 19-24 rapidly. Other recent efforts, including the work described in this master thesis, have gathered and analyzed data from WhatsApp chat groups, focusing on textual interactions in these groups. For example, the studies in [18, 11] also analyzed WhatsApp groups behavior with data collected by the same process of [30], as we also did in this work. Melo *et al.* studied the anatomy of WhatsApp groups of 3 countries analyzing how the forwarding tools contribute to the virality of misinformation and whether system limitations are capable of preventing the spread of content. The authors concluded that those limits are not effective in preventing a message to reach the entire network quickly [18]. Caetano *et al.*, in turn, presented a large-scale study of collective user attention on WhatsApp political and non-political groups [11]. The authors proposed to study attention by applying a cascade framework. They found that cascades with false information in political groups tend to be deeper, broader and reach more users than the same type of cascades in non-political groups.

Our present effort provides a deeper understanding of the content exchanged in WhatsApp, unveiling, among other findings, the spread of misinformation campaigns through images and textual messages in the platform during the Brazilian presidential elections in 2019. In sum, while prior studies provide valuable knowledge about WhatsApp as an emerging social network and information dissemination vehicle, the analysis of information spread in the system is still at a very early stage. This master thesis greatly adds to the current literature by focusing on properties and propagation dynamics of information on WhatsApp and how its propagation crosses the boundaries to and from the Web. To our knowledge, this is one of the first efforts to perform such analysis, focusing on textual messages and images shared in the groups.

## Chapter 3

# WhatsApp Data Collection

In this chapter, we describe how we built the dataset of WhatsApp public groups and the methods we used to process and analyze the collected data. Then, we present an overview of the content present in our dataset.

### 3.1 Methodology

Figure 3.1, presents a flowchart with the steps of our WhatsApp data collection methodology. As a first step of our data collection, we had to identify a considerable number of publicly accessible groups. To that end, we used the URL pattern "*chat.whatsapp.com*", which is commonly used in invitations to join WhatsApp groups, as a search query and submitted it to Google, Twitter, and Facebook search engines. We restricted our search space to groups related to Brazilian politics, by including in each search query a word from a dictionary related to the 2018 Brazilian elections (see Appendix A).<sup>1</sup> This dictionary contains the names of politicians, political parties, as well as words associated with political extremism. The vast majority of these publicly accessible groups links originated from specific websites that share WhatsApp groups as *grupowhats.online*<sup>2</sup>. Others came from political groups or communities or candidates posted on social media like Facebook and Twitter. Finally, we performed a manual inspection of the collected group names to filter out those unrelated to politics. In total, we found 3,444 distinct links for publicly accessible groups, out of which only 1,828 were valid (i.e., unbroken), identified using a script developed by Garimella *et al.* [30].

As a second step, we selected a number of valid groups to monitor. This monitoring involves joining each group using a cell phone. Thus, the number of groups (see Section 3.2) monitored was constrained by the available devices and their resources (memory). We joined each selected group using our available cell phones, 2 Android and 1 IOS devices,

---

<sup>1</sup><https://goo.gl/PdwAfV>

<sup>2</sup><http://grupowhats.online/>



Figure 3.1: WhatsApp Data Collection Flowchart

and a tool developed by Garimella *et al.* [30] to automatically join groups. We then periodically downloaded all data shared in each group and stored them in a database, using the *WebWhatsAppAPI*<sup>3</sup>. This tool provides an interface in Python to send and receive messages by *WhatsApp Web* and uses *Selenium* to automate the application through the browser. Specifically, stored data can be grouped into: images, videos, audio messages, external links, and text messages. From each message, we extracted its group name (i.e., the group the message was posted), a group ID, a user ID, and timestamp. That is, we mapped telephone numbers and user names into unique user identifiers<sup>4</sup>, discarding the original information afterwards. For the media messages, we also downloaded their respective files and used their filenames as a reference to the message. All collected data were stored in a Linux server where we run the *Web WhatsApp API*, we also deleted all messages from the 3 cell phones periodically.

To our knowledge, this work is the first effort that aims to explore the political debate in publicly accessible WhatsApp groups. Also, it proposes a methodology to infer which identified publicly accessible groups are related to politics. Unfortunately, we are not aware of an approach that would allow us to assess the representativeness of our data as even the total number of groups available in the country is not of public knowledge. We emphasize, though, that all sensitive information (e.g., user names and phone numbers) were not stored in our dataset.

## 3.2 Dataset

Our data collection focuses on the period of the first round of 2018 Brazilian presidential elections campaign (August 16<sup>th</sup> to October 7<sup>th</sup> 2018). During this period, according to the Brazilian election law, political parties and coalitions are authorized until October 6, the day before the first shift, to campaign on paid advertisements in the

<sup>3</sup>WhatsAppAPI available on: <http://github.com/mukulhase/WebWhatsapp-Wrapper>

<sup>4</sup>Throughout this master thesis we refer to such identifiers as users. Yet, they are indeed unique telephone numbers, as we are not able to identify multiple devices of the same user.

Table 3.1: Overview of our dataset.

	Election Campaign
#Groups	364
#Total Users	18,725
#Total Messages	789,914
#Textual Messages	591,162
#Images	110,954
#Videos	73,310
#Audios	14,488
#URLs	92,654
Filtering: Distinct Images	
#Distinct Images	69,685
Filtering: textual messages with > 180 characters	
#Textual Messages	59,979
#Distinct Textual Messages	37,674
#URLs	19,502

print media and on the Internet. We focus on this period to understand, the presidential elections campaign on publicly accessible politically oriented WhatsApp Groups. We note, however, that we found more right-wing groups in our methodology for searching public groups with political themes and that our database has a bias in this regard.

Table 3.1 provides an overview of our dataset, showing the total number of messages shared as well as the number of messages per type of content (text<sup>5</sup>, image, video and audios). Note that most shared messages are indeed textual content with 74% of all content, and images are the most frequent type of media content in the dataset, reaching roughly 15% of all content shared in the monitored groups during the election period. Note also a large number of links to websites (last row) present in the text messages.

### 3.3 Data Filtering

Images and textual messages are the most frequent type of content in our collected data, and they are thus the focus of our analyses. However, in order to analyze the spread of a particular content, it is necessary to identify duplicates, i.e., messages containing the same content. In the following we discuss how we perform such identification for images and textual messages.

<sup>5</sup>Only messages that are entirely composed of textual content are counted as text messages.



### 3.3.1 Images

Figure 3.2, presents a flowchart with the steps of our methodology for filtering images and removing duplicate images of our dataset. In order to identify duplicates of the same image, we used the *Perceptual Hashing (pHash)* algorithm [64] to calculate a fingerprint for each image. We were then able to group images having the same hash-values based on human eye perception as duplicates. As presented in Table 3.1 (lower portion), this filtering left us with 69,685 unique images during the election period. We selected a representative image for each content, keeping information about the groups each image was sent to and their timestamps.

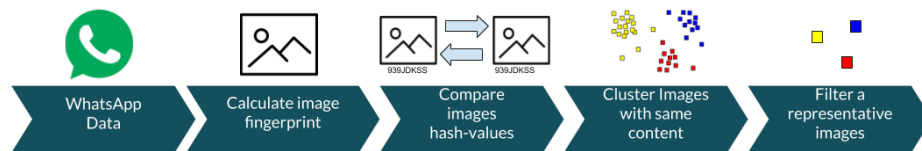


Figure 3.2: Images Filtering Flowchart

As we here aim to explore the presence of WhatsApp content on the Web, we delved further into the images shared on the monitored groups and developed a tool to collect Web pages in which those images have also appeared. The tool exploits the capability of searching for images provided by Google search, where a user can submit an image as query, and obtain as result webpages that include matching images along with their post dates. As will be discussed in Chapter 4, this information allows us to analyze temporal sharing patterns such as the time interval between the first appearance of an image on WhatsApp and elsewhere on the Web.

In order to explore the content veracity of the images shared in the groups, we extended the tool to automatically identify whether each image was fact-checked by some of the main Brazilian fact-checking agencies. This is done by checking whether fact-checking websites appear in the results of the initial Google Image search. If it does, then we proceed to parse the fact-checking web page and retrieve the fact-checker verdict for the image. So, we also added information in our dataset about where and when each image from WhatsApp appeared on the web, and also if the image content was checked with misinformation.

### 3.3.2 Textual Messages

Since we aim at studying the presence of misinformation, we analyzed only messages with at least 180 characters, which is a satisfactory minimum size for texts that bring some information or summary of news, to avoid small talks and greetings. This filtering left us with 59,979 textual messages, many of which contain one or more URLs to websites and external news, summing up almost 19,502 links in our analyzed messages.



Figure 3.3: Images Filtering Flowchart

Figure 3.3, presents a flowchart with the steps of our methodology for filtering textual messages and removing duplicates of our dataset. We grouped similar content by computing the Jaccard similarity [41] between pairs of messages. The Jaccard similarity between messages  $m_i$  and  $m_j$  is computed as the ratio of the number of common words in both  $m_i$  and  $m_j$  to the number of words in the union of both messages. Messages with a similarity greater than 0.7 were considered the same and being thus grouped and considered as (semi-)duplicates. The choice of the threshold was made empirically, once. After manually inspected a sample of the messages, we note several messages that carried the same information despite differences in the use of words and emotions, and no content difference was found in pairs of messages with a similarity greater than 0.7. In this process, a representative message for each content was randomly selected, keeping information about the groups each content was sent to and their timestamps. In total, we identified 37,674 distinct textual messages (62% of the total textual messages) in the monitored groups during the election period.

## 3.4 Overview

In this section, we provide a brief overview of the type of content present in our dataset, notably textual messages and media content (audios, videos, and images). We also briefly discuss the presence of URLs in textual messages. The discussion is based on

Table 3.1 as well as on Figure 3.4, which shows the complementary cumulative distribution of the number of messages of different media types shared per day, across all monitored groups.

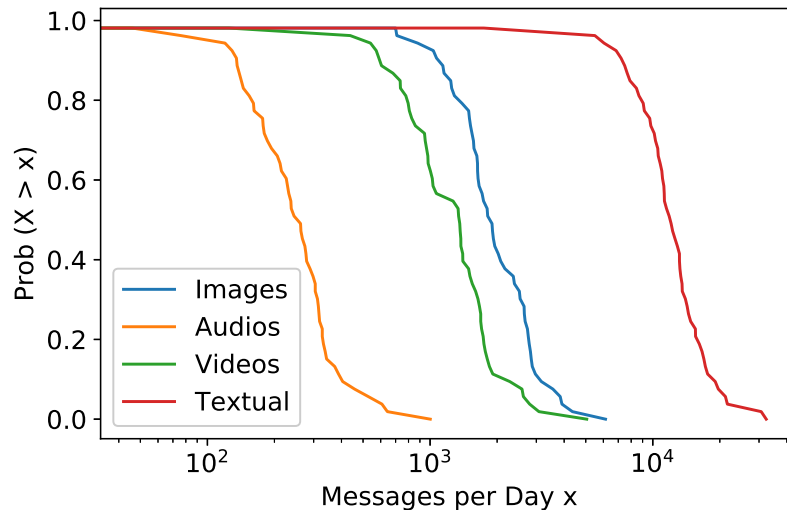


Figure 3.4: Numbers of daily messages with media content shared on all groups

### Textual Messages

Textual messages represent the most frequent type of message shared on the monitored groups, with a total of 591,162 (almost 75%) textual messages shared during the election campaign periods. Figure 3.4 shows that up to 9,000 textual messages were shared in 60% of the monitored days. In about 10% of the days, the number of textual messages shared on the groups on a single day exceeded 19,000 in this period.

### URLs and Webpage domains

As shown in Table 3.1, a total of 92,654 URLs were shared (as part of textual messages) in the monitored groups during the election campaign. They correspond to 9% of the total amount of messages we gathered for this period. The nature of these URLs varies from links to news websites, blogs, entertainment, and other social networks to even links to other WhatsApp groups. Also, 45% of all URLs shared during the election campaign period are unique, indicating less diversity and more repetition of the links during that period.

### Audios

Our datasets contain 14,488 audio messages, which correspond to 2% of all messages gathered during the monitored period. As shown in Figure 3.4, audio is the least frequent media type shared in the monitored groups. For example, up to 500 audios were shared in 60% of the monitored days. Note however that, this number reached a peak two days before the election day, with 1,002 audio messages shared on all monitored groups.

### Videos

In total, 73,310 videos were shared during the election campaign period, which corresponds to 9% of all messages shared during the period. Also, according to Figure 3.4, approximately 1,300 videos were shared daily in 60% of the days (with a maximum of 5,052 videos on a single day). Those numbers show that videos are also very popular in the platform.

### **Images**

Finally, Figure 3.4 shows that images represent the most popular media content shared on the monitored groups, with a total of 110,954 images (almost 15% of all images) shared during the election campaign period. According to the figure, this higher frequency happens on a daily basis. Up to 2,000 images were shared in 60% of the monitored days. Even more, in about 5% of the days, the number of images shared on the groups on a single day exceeded 6,100 images (these peaks occurred in the week before the election day).

Given the popularity of images and textual messages, we select these two subsets of messages to delve further and we do not go deeper into the study of audios and videos. In the next two chapters we present a thorough characterization of the content properties and propagation dynamics of each type of message, image and textual content, contrasting messages containing previously checked misinformation with the rest. We also present a first analysis of the network structure of WhatsApp groups.

## Chapter 4

# Image and Textual Content Properties

In this chapter, we characterize the content properties of WhatsApp messages focusing on images (Section 4.1) and textual messages (Section 4.2) (RQ1). We compare messages containing misinformation with the rest, here referred to as unchecked, aiming at highlighting properties that distinguish them (RQ2). To distinguish between them we refer to the former as misinformation and to the latter as unchecked, since the veracity of their content could be not necessarily checked. We cannot guarantee the absence of misinformation in the unchecked messages, given that such an assertion is restricted by the availability of checked facts. Yet, we expect that we were able to catch most messages containing misinformation in our dataset, especially those with greater impact on users, as they most probably were identified by the fact-checkers.

### 4.1 Images

In this section, we start by contrasting images sent in WhatsApp groups containing previously reported misinformation with the other unchecked images. We first characterize the content of WhatsApp images, their main topics, as well as the domains on the Web and Twitter accounts where those images also appeared. Finally, we analyze how images shared on WhatsApp groups propagate to and from the Web.

We start by presenting some key measures related to the sharing of images within the monitored groups. Table 4.1 presents averages, standard deviations, and maximum values of the numbers of images shared by each user and within each monitored group as well as number of users sharing images in each group and the total number of times each image was shared (across all groups). Note that most images are shared only a few times (once, on average), as only a few images are widely shared. Interestingly, we found that

the groups with the largest number of images shared during the election campaign<sup>1</sup> are indeed the groups with the largest numbers of users sharing this type of content (with 211 users).

Table 4.1: Sharing of images on monitored WhatsApp groups.

	Mean	Standard Deviation	Maximum
Number of images per Group	345	580.531	4,320
Number of users per Group	34	41.347	211
Number of images per User	10	32.322	1,612
Number of shares for Image (total)	1	4.512	125

### 4.1.1 Misinformation

We start by looking at the presence of misinformation in the images shared on WhatsApp groups. First, we discuss two techniques used to identify misinformation in the images in our datasets. We analyze their characteristics comparing these images with the rest of our WhatsApp data.

#### 4.1.1.1 Labeling with a Fact-Checking Agency

We created a list of the most shared images during the election campaign period and gave them to one of the most important fact-checking agencies in Brazil, *Lupa*<sup>2</sup>. They checked the veracity of each of these images following a methodology similar to other fact-checking agencies around the world (e.g., the American *Politifact*<sup>3</sup> and the Argentinian *Chequeado*<sup>4</sup>). They analyzed whether these images contained factual information as opposed to opinions since it is not possible to check the latter. Out of a total of 61 of the most shared images during the election campaign period, 47 were marked as factual. Out of these factual images, they found that 22 had already been checked by other fact-checking agencies: 17 images had been checked as containing misinformation, and only 5 images had been checked as true. These results show an expressive number

<sup>1</sup>The group with the largest number of shared images was "BOLSONARO PRESIDENTE", with 4,320 images

<sup>2</sup><https://piaui.folha.uol.com.br/lupa>

<sup>3</sup><https://www.politifact.com>

<sup>4</sup><https://chequeado.com>

of images with misinformation that went viral in WhatsApp during the 2018 Brazilian elections. In terms of percentages, 36.2% of the images with factual information were checked as containing misinformation, whereas 53.2% of them include misleading and inconclusive content (not supported by public information), and only 10.6% were verified as true. Examples of images checked as misinformation are shown in Figure 4.1.

Figure 4.1(a) is an edited image of the Brazilian former president Dilma Rousseff, who was impeached in 2016 [81], alongside Fidel Castro, former president of Cuba. At the time this picture of Castro was taken, Dilma was 11 years old. Thus, the image is clearly fake. It was the most popular image in the monitored groups during the analyzed period. Figure 4.1(b) is an edited image of the former Brazilian president Lula, imprisoned for corruption at the time of monitoring [16, 101], meeting the aggressor responsible for stabbing the then presidential candidate Jair Bolsonaro during a campaign rally [67]. The intention of the image was to associate Lula with the attack against Bolsonaro.



(a) False edited photo of ex-president Dilma next to Fidel Castro



(b) Fake photo of Bolsonaro's aggressor next to Lula

Figure 4.1: Images checked as fake by both fact-checking methodologies.

#### 4.1.1.2 An Automatic Methodology for Finding Misinformation

Recently, Facebook has announced partnerships with many third-party fact-checking organizations, through which Facebook demote or reduce the visibility of links rated as false [56]. This kind of partnership neglects misinformation in images, as fact-checkers only provide rates to links containing stories with misinformation. Next, we provide a strategy to connect the false stories found in images shared with external links that ap-

pear on the Web, providing a simple way for Facebook to demote links containing images with misinformation identified on WhatsApp.



Figure 4.2: Flowchart of an Automatic Methodology for Finding Misinformation in Images

Figure 4.2, presents a flowchart with the steps of our methodology for finding misinformation in images. First, we identified the main fact-checking agencies in Brazil<sup>5</sup>. We then automatized the process creating a script that searches each image that was shared more than once on the WhatsApp groups on the Web by using the Google Image search as explained in Subsection 3.3.1. From (13688) images that were shared more than once, we obtained results for 10% of the images in this set. Given the search results for an image, we checked whether any of the returned pages belong to one of the fact-checking domains. If so, we parsed the fact-checking page and automatically labeled the image as fake or true depending on how the image was tagged on the fact-checking page. If at least one fact-checking site tagged the image’s information as fake, we also labeled it as fake.

We applied this methodology to all images (12319) from our dataset that was shared more than once and also appeared in a domain in the Web and found 70 images containing misinformation. We compared the 70 images with misinformation identified by the automatic process with the 17 images checked as fake by Lupa (see previous section), obtaining an overlap of only 2 images. Thus we built a single dataset of 85 images with misinformation identified by official fact-checking agencies.

Table 4.2 presents a comparison of the images with misinformation and with unchecked content shared during the election campaign, showing the numbers of distinct images, users who shared those images, groups in which those images were shared and the total number of shares. Note that, even though the number of distinct images with misinformation is small (85), these images summed up 1,168 shares posted by 624 different users in 157 different groups. Despite representing less than 1% of all images shared, these images appeared in 44% of the monitored groups in the period of the election campaign, effectively reaching a large user population. Also, note that nearly 5.7% of all users shared images with misinformation.

In the following, we analyze the images in this dataset, focusing on their content and other websites where they also appear. We also compare properties of these images with those of the other images shared during the election campaign period.

<sup>5</sup>Fact-checking agencies: *Boatos.org*: <https://www.boatos.org>; *e-Farsas*: <http://www.e-farsas.com>; *Comprova*: <https://projeto comprova.com.br>; *Lupa*: <https://piaui.folha.uol.com.br/lupa>;



Table 4.2: Overview of images shared during election campaign period: misinformation versus unchecked content.

	Misinformation	Unchecked
Number of groups in which images were shared	157	351
Number of users who shared images	624	10,339
Number of unique images	85	69,590
Number of total shares of images	1,168	109,791

### 4.1.2 Content Labeling

A WhatsApp group is usually meant to be a space for discussions about a specific subject such as politics, education, games. However, the content shared itself may diverge from the group subject given the will of their participants. For example, as in other Web systems, WhatsApp groups are susceptible to spam activity (e.g. advertisements or inappropriate content). To understand the kinds of images shared on our selected groups, we first categorize the images by performing content labeling and analyze the distribution of images across categories. We also discuss the appearance of the same images on other websites and social networks.

We asked three volunteers to label a sample of the most shared images during the election campaign period. The sample contains the top-100 most shared images, considering the whole monitored period. For the images with misinformation, the sample contains all 85 images with misinformation identified by the fact-checking agencies.

A taxonomy guideline document with instructions was given to the volunteers with the following directions: (i) observe an image and, read the text on it, if available; (ii) if there is a text, check the existence of any citation to a website or other source; (iii) check if the following content types are present in this post: *Political Content*; *News*; *Opinion*; *Satire*; *Activism*; (iv) identify possible inappropriate, offensive or even illegal content by checking for the presence of *Dissemination of Hate*; *Violence*; or *Promotion of Illicit Products as Inappropriate Content*; (v) you can classify a post with more than one category (e.g., News and Political Content) or none of them; and finally (vi) if you cannot fit the image in any of the listed categories or are unable to establish its category, label the image as *Others*. Table 4.3 lists the categories used to label the sampled images.

After each of the three volunteers annotated each image according to the categories in Table 4.3, we measured the inter-annotator agreement in terms of the *Fleiss's*  $\kappa$  [26]. We assumed that consensus was reached if the null hypothesis of negative or no agreement  $\kappa = 0$  can be rejected. Since the same image may fit more than one category, we applied the test individually for each category, averaging the  $\kappa$  scores obtained, as we can see in

---

*Globo G1*: <http://g1.globo.com/fato-ou-fake>; and *Aos Fatos*: <https://aosfatos.org>.

Category	Description
Political	Information about a candidate or party
News	News information with a quote.
Satire	humorous content regarding current events.
Inappropriate	Illicit products, violence, hate speech or pornographic content
Activism	Popular movements and protests
Opinion	Which expresses a personal opinion or comment.
Others	Does not fit into any other category

Table 4.3: Image Categories

Table 4.4. The category with the higher agreement was News and the lower was Opinion. Overall, we obtained moderate agreement among the annotators, with average  $\kappa$  equal to 0.39 for the sample of images analyzed. This result is reasonable given that some categories are very broad and distinctions are somewhat blurred. In the following, we assume that an image belongs to a category if at least two of the annotators agreed upon that category.

Category	Kappa
Political	0.30
News	0.66
Satire	0.35
Inappropriate	0.44
Activism	0.40
Opinion	0.19

Table 4.4: Image Categories *Fleiss's*  $\kappa$ 

Figure 4.3 shows the distributions of the image categories in each sample. As expected, most images are related to politics (76% for images with unchecked content and 74% for images with misinformation). The majority of the images with misinformation were hoaxes, used to misrepresent well-known politicians and candidates and personalities involved with the elections. Images with misinformation have also less expressive topics like false news and a lot of activism scenarios deceived. We also observed false images with inappropriate content with the intent of harming the opposition group, or candidates in the election scenario. The presence of false opinions from personalities in the images are also worth noting: we observed a lot of prints of social networks profiles supporting candidates or groups, corresponding to 11% of all images with misinformation. In contrast, images with opinions from personalities were much more popular in the sample of images with unchecked content, corresponding to 30% of all images in our sample. News and images with activism are also frequent among the unchecked content, but no image in our sample contains explicit advertisements.

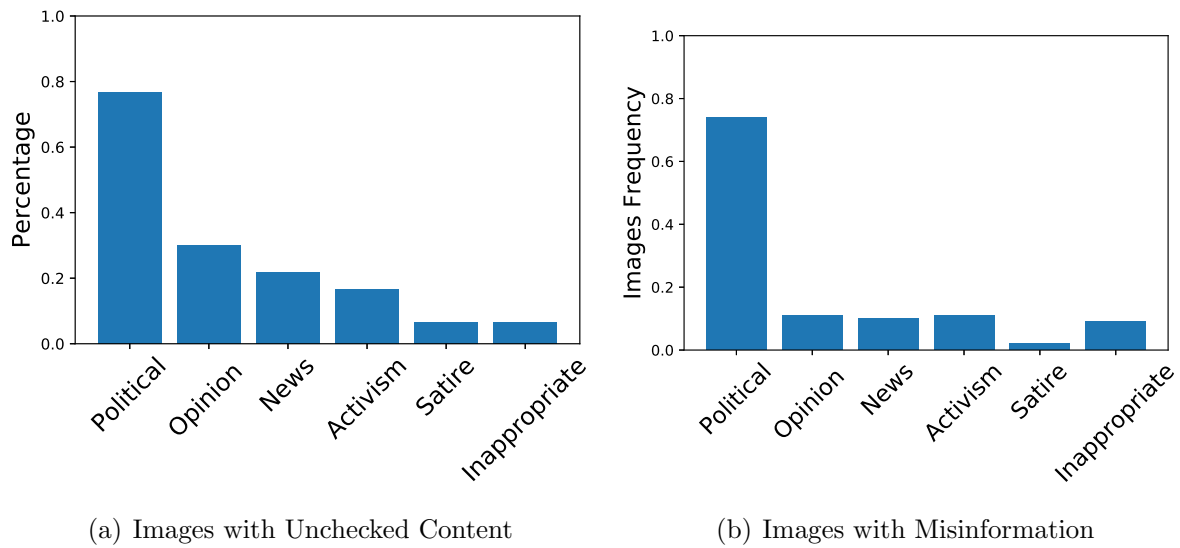


Figure 4.3: Distributions of image categories.

### 4.1.3 WhatsApp Images in other Websites

We now analyze the extent to which the images shared on our monitored WhatsApp groups have also appeared on other websites including social networks and blogs. We do so by searching for the observed images using the Google Images search engine, as discussed in Section 3.1.

Figure 4.4 shows the most popular domains returned by Google Images for the images with unchecked content and for the set with misinformation. Notice that online social networks like Twitter, Facebook and, Google+ are among the most frequent domains where the images were posted for both sets. TrendsMap, a website that shows visualizations of the trends on Twitter, was also popular. Similarly, image apps like Deskgram as well as Blogspot are popular domains, especially the latter, suggesting that a large fraction of image content shared on WhatsApp groups may indeed have blogs as possible sources.

Compared to unchecked content, images with misinformation appeared much more frequently on other social networks: as examples, while the fractions of images with misinformation that also appeared on Twitter and Facebook reach 86% and 55%, respectively, corresponding fractions for images with unchecked content are only 36% and 20%, respectively. We note that Twitter may be highly frequent, because it is a more open social network. Those results, however, suggest that images with misinformation may have a viral behavior beyond WhatsApp. Moreover, other social networks like Google+ and Youtube as well as blogs (e.g., Blogspot) also quite a large presence among the images with misinformation. Note the presence of UOL, Boatos and Globo among the domains

where images with misinformation often appeared. Most probably those appearances refer to efforts of checking and reporting fake news.

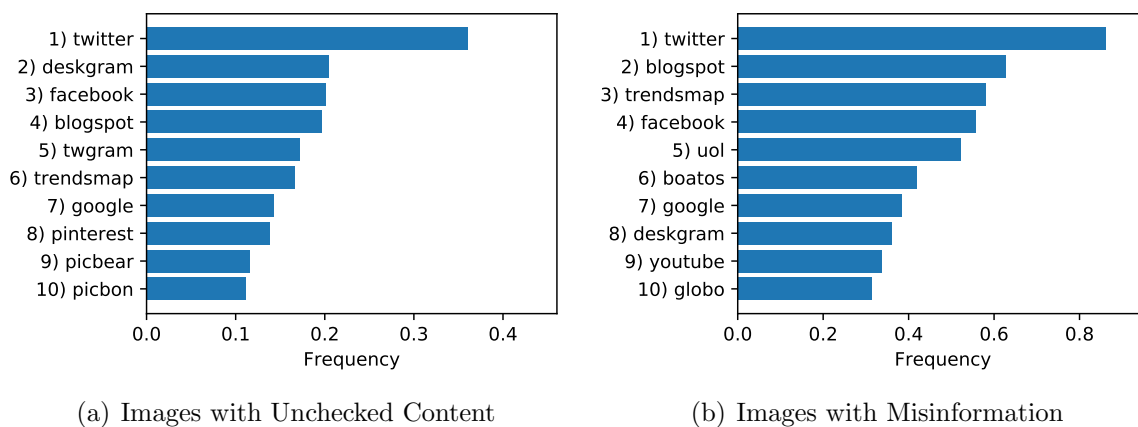


Figure 4.4: Most popular domains for images shared on WhatsApp publicly accessible groups.

To further analyze the images that were also posted on Twitter, Figure 4.5 shows the Twitter accounts that most often shared both images with misinformation and the others, the figure shows the number of posts by each account containing images of each set. For both images with misinformation and with unchecked content, we can observe the presence of some official journalistic accounts (*folha*, *agencialupa* and the journalist *blogdonoblat*) and official accounts of the presidential candidate Fernando Haddad and his vice, Manuela d’Avila. We note these profiles posted images containing misinformation most probably with the purpose of repudiating them, acting as fact-checking accounts. Yet, some other accounts acted as misinformation broadcasters by spreading it further through the network.

In Figure 4.5(a), there are also more of the main Brazilian news official accounts (*folha*, *estadao* and *g1*) and official accounts of politicians, like the presidential candidates Jair Bolsonaro and Fernando Haddad and his vice, Manuela d’Avila and candidates for deputies like Carlos Bolsonaro. The great number of Twitter accounts related to politics go according to the theme of the monitored groups, however, the large number of official news accounts indicate that images from news are frequent in groups on WhatsApp.

## 4.2 Textual Messages

In this section, we shift our focus to textual messages, and analyze the properties of messages containing misinformation and unchecked content, highlighting differences

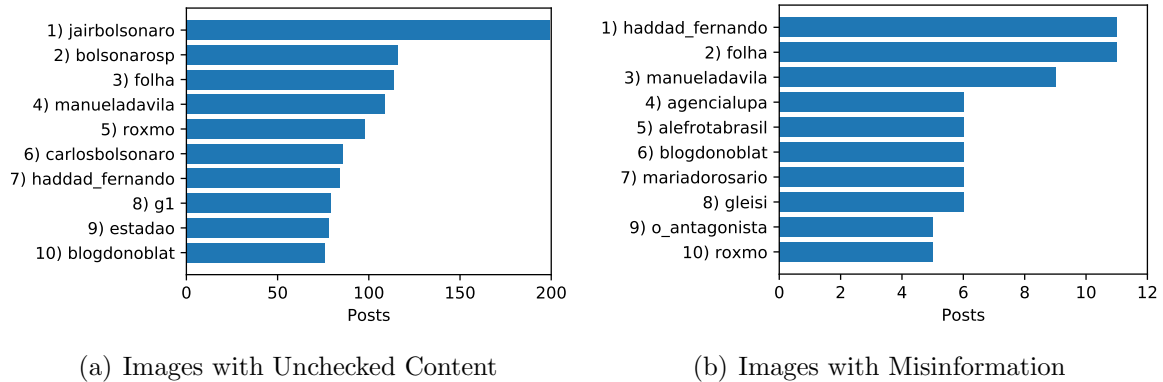


Figure 4.5: Most popular Twitter accounts for images shared on WhatsApp publicly accessible groups.

between them. Before discussing our results, we first present how we identified textual messages with misinformation.

### 4.2.1 Identifying Misinformation

To identify misinformation in our textual messages, we collected facts there were previously checked as *fake* by fact-checking websites and compared them to the messages in our filtered dataset (i.e., the dataset obtained after filtering out textual messages with less than 180 characters). Specifically, we crawled checked information (news or claims) from the same six Brazilian fact-checking sites we used to identify misinformation in the images (see Section 4.1) and we parsed the HTML of each fact-checking site to crawl the checked information. Figure 4.6, presents a flowchart with the steps of our methodology for finding misinformation in textual messages. We collected all posted facts published during the year of 2018, including title (or claim), URL, description, summary, associated images (links, if available), authors (if available), date, and label (i.e. fake or not). In total, 1,234 facts labeled as fake were collected.



Figure 4.6: Flowchart of Methodology for Finding Misinformation in Textual Messages

We then computed the text similarity between each textual message in our What-

sApp dataset and each collected fact labeled as fake by at least one of the fact-checking websites. For the latter, we experimented with using only the contents of the summary field and using the description, which contains a more detailed presentation of the fact. We found that using the summary leads to more accurate textual matching possibly because WhatsApp messages tend to be more direct to the point. We first pre-processed each piece of textual content (WhatsApp message and fact summary) using a version of the Spacy natural language processing toolkit specific to Portuguese<sup>6</sup> to remove stop words and accents as well as stemming words. Each piece of content was then modeled as a bag of words, by means of a TF-IDF vectorial representation, widely used in information retrieval [65]. Given a WhatsApp message  $m$  and a fact summary  $s$ , represented by their TF-IDF vectors  $v_m$  and  $v_s$ , respectively, we computed their textual similarity by means of the cosine similarity, defined as  $\cos(v_m, v_s) = \frac{v_m \cdot v_s}{\|v_m\| \|v_s\|}$ .<sup>7</sup>

We computed the similarity scores between all pairs of messages and fake fact summaries. Note that the two pieces of content may refer to the same fact and yet have (cosine) similarity below the maximum of 1. Thus, the identification of misinformation depends on some similarity threshold. To define such threshold, we first manually compared a sample of 100 WhatsApp messages with the fact summaries, determining whenever both referred to the same (fake) fact. We then compared this manual label with the similarity cosine scores. No match was found in our manual labeling between contents with cosine score below 0.4. Thus, any WhatsApp message whose cosine similarity with any of the fake fact summaries was above 0.4 was considered *suspicious of carrying misinformation*. All suspicious messages were then manually analyzed and compared against the fact summaries. We note that messages with high similarity scores, but containing retractions (e.g., links to fact-checking websites refuting the original content) were manually excluded from the misinformation dataset. This process led to the identification of 69 *distinct* textual messages containing previously checked misinformation. These messages were shared 578 times in our dataset<sup>8</sup>. We found on textual messages a slightly smaller number of messages containing previously checked misinformation than was found in the images, however, images represent a greater volume of messages in our dataset.

In the following sections, section, we compare the textual properties of messages containing misinformation with the other textual messages in our dataset (unchecked content).

---

<sup>6</sup><https://spacy.io/>

<sup>7</sup>We did experiment with other similarity metrics, notably WMD (*Word Mover's Distance*)[51], which covers the semantics of sentences, and the results were similar, but with a higher processing cost.

<sup>8</sup>Throughout this master thesis we use the term *sharing* of a message as a synonym of posting a message in a WhatsApp group. In that sense, the same message (same content) may be shared/posted multiple times by one or more users.

## 4.2.2 Textual Properties

In this section, we analyze textual properties of WhatsApp messages containing misinformation as well as unchecked content, highlighting differences between them. Our analyses cover message size, psychological linguistic features, sentiment analysis as well as the main topics and frequent words present in each type of message.

### 4.2.2.1 Message Sizes

We start by looking at the sizes of the messages. Figures 4.7(a) and 4.7(b) show the cumulative distributions function (CDF) of the numbers of words and characters in the messages containing misinformation as well as in the messages whose content was unchecked. We compared the two distributions using the Kolmogorov-Smirnov test [59] with 95% confidence level, with the null hypothesis that two samples have the same distribution and we found a slightly statistical difference ( $p$ -value of 0.059). According to Figure 4.7(a), 20% of the messages with misinformation have up to approximately 15 words. Those are often messages with links to websites or blogs publicizing fake news. In contrast, the same fraction of unchecked messages has up to 20 words. Indeed, considering only messages of intermediate size (up to 50 words), those carrying misinformation tends to be shorter. The two distributions continue very similar up to roughly 748 words, which is the maximum number of words in all messages with misinformation analyzed. Yet, there are a few longer messages (more than 5,000 words) with unchecked content in the dataset. In general, despite the variability in intermediate sizes, messages with misinformation tend to have fewer words<sup>9</sup>.

Figure 4.7(b) shows that both distributions of numbers of characters are very similar up to around 4,000 characters, with the Kolmogorov-Smirnov test we found no statistical difference ( $p$ -value of 0.38). Roughly 60% of both types of messages have up to 280 characters, and the medium size is 459 and 472 characters for messages with misinformation and unchecked content, respectively. However, once again, we do find some very long messages (up to 61,681 characters) among those with unchecked content.

---

<sup>9</sup>We note that the larger presence of links in messages with misinformation, as will be discussed in the next section, does not impact the difference in length as each link is counted as one word.

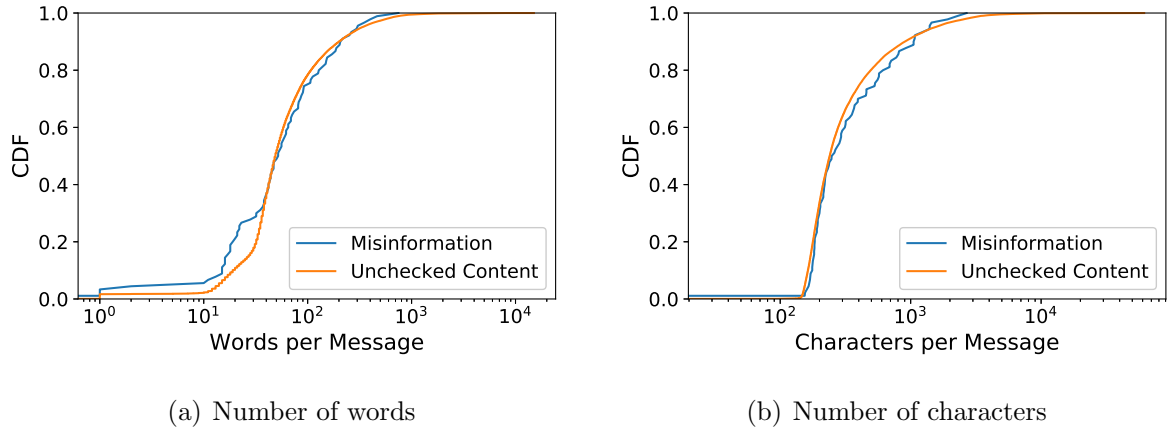


Figure 4.7: Distributions of message sizes

#### 4.2.2.2 Psychological Linguistic Features

Textual messages with misinformation may contain psychological and cognitive elements that can trigger specific reactions, possibly boosting the sharing of the message to others. In order to study the distribution of psycholinguistic elements in the textual messages, we extracted these types of features from the texts using the 2015 version of the Linguistic Inquiry and Word Count (LIWC) [96]. LIWC is a psycholinguistic lexicon system that categorizes words into psychologically meaningful groups. We used the dictionary for the Portuguese language, which is organized as a hierarchy of categories and subcategories, all of which form the set of LIWC attributes. Examples include linguistic style attributes, affective attributes, and cognitive attributes. Positive emotions, negative emotions, anxiety, anger are examples of subcategories of the affective attributes, whereas insight, causation, discrepancy are examples of subcategories of the cognitive attributes. In total, there are 92 LIWC attributes. Each such (sub)category is characterized by a set of words from the dictionary. Examples of words representing the anger attribute in the LIWC Portuguese dictionary are *hate*, *kill*, *pissed* (translated to English). Given an input text, we compute the value of a LIWC attribute as the percentage of words in the text that represent the given attribute. Note that, as such, an attribute value is normalized to the size of each message.

We characterized both messages with misinformation and unchecked content with respect to the presence of psycholinguistic elements by computing the distributions of attribute values for each LIWC attribute for both sets of messages. As a first step to narrow our attention to the most distinguishing attributes, we compared both distributions using a Kolmogorov-Smirnov (KS) test [59], which is a non-parametric test of equality of continuous distributions, in which the null hypothesis states that the two input samples



have the same distribution.

Messages with misinformation have a larger presence of URLs: 50% of the messages with misinformation have at least one URL, whereas only 32% of the other messages contain such links. The presence of such URLs emphasizes the linguistic features related to punctuations that are frequent in links. Thus, in order to investigate the presence of other psycholinguistic features, we removed the links from all messages in this analysis.

We identified 7 (out of 92) attributes for which, according to the KS test, the two distributions differ with a confidence level of 95%. We then computed the relative difference between the *average* values of each such attribute for messages with misinformation and messages with unchecked content. These differences are shown in Figure 4.8. As shown in the figure, the attributes with a significantly greater presence in messages with misinformation are subcategories of the linguistic attributes (*we*, *they*, *present*, *exclaim*) and psychological attributes (*insight*, *inhibition*, *sexual*).

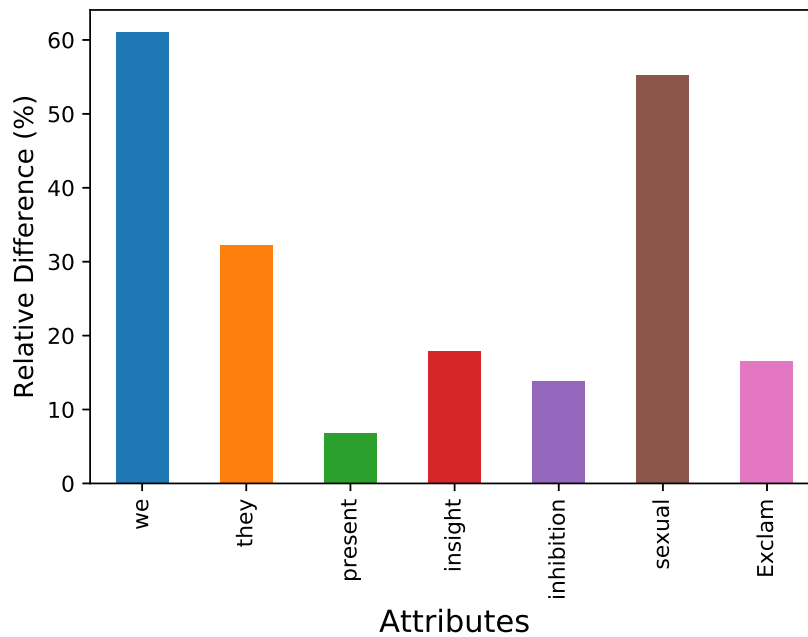


Figure 4.8: LIWC attributes that occur more frequently in messages with misinformation.

We identified some significant presence of the attributes *we* and *they*, representing words and verbs in the first and third-person plural respectively. The former was used in phrases aiming at aggregating the community towards the same goal, and the latter to refer to third parties. The attribute *present* indicate frequent use of verbs in the present tense and misinformation in current news and events. The exclamation mark was also observed with the attribute *Exclaim*, used in messages with misinformation content, to drive the attention and appeal to a more emotive speech.

The *insight* attribute is a cognitive process characterized by words like *attention*, *warning*, *look*, and *listen*, which occurred very often in messages with misinformation,

especially those structured as chain messages, where warnings and verbs in the imperative are common. We also noticed that messages with those words were shared 40% more times than the remaining, on average. Rumors about voter turnout and denial of previously reported facts were also observed in the messages with misinformation in our dataset, with a larger presence of words like *deny*, *null*, and *block* which characterize the *inhibition* attribute. The *sexual* attribute is represented by words such as *virgin*, *orgy* and *nudism*, often related to offensive content. We conjecture that the somewhat higher frequency of such attribute in messages with misinformation is due to the presence of false stories and hate speech content towards some political opposition groups. We also observed some sensationalist headlines that use sexual content to attract attention.

#### 4.2.2.3 Sentiment Analysis

Sentiment analysis has become an extremely popular tool to capture text polarity, especially in social media data [62]. In order to investigate the overall subjective cues of sentiment in the WhatsApp textual messages, we used a Portuguese version<sup>10</sup> of SentiStrength method [98] to measure the polarity of each piece of content. SentiStrength is a well-established method that implements a combination of supervised learning techniques with a set of rules that impact the "strength" of the opinion contained in the message. This technique has already been applied in several domains (e.g., to capture the strength of sentiments expressed in headlines of online news [72]). We here employ it to investigate whether there are differences in the sentiment of messages carrying misinformation when compared to the rest.

Figure 4.9 shows the percentages of positive, neutral and negative messages with misinformation and carrying unchecked content. It is interesting to note the very large volume of negative messages in both groups. A large presence of negative content has also been previously reported for Twitter [97]. However, the results in Figure 4.9 suggest an even stronger bias towards a negative discourse on WhatsApp. Moreover, there are more positive messages than neutral ones (also in both groups), which evidences the polarized nature of the data, leaning more often towards more extreme feelings rather than neutral text.

Comparing messages with misinformation with those with unchecked content, we do observe some differences, but they are small. In particular, messages with misinformation are slightly more negative. Such difference is indeed statistically significant by a Kruskal-Wallis H-test [48], with p-value < 0.005. This finding is in agreement with previ-

<sup>10</sup>Available in: [sentistrength.wlv.ac.uk](http://sentistrength.wlv.ac.uk).

ous observations that misinformation content tends to be more negative [115], especially within polarized communities. Moreover, inspired by previous results on online news [72], one could speculate that messages with misinformation tend to be more negative as a mechanism to attract readers. As we will see in Section 5.1, such messages are indeed shared a larger number of times in our dataset.

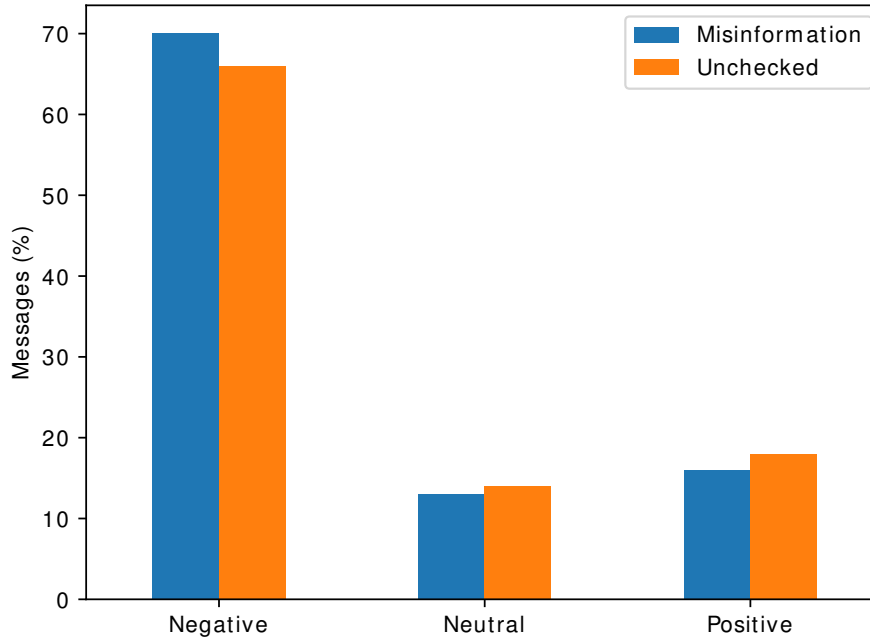


Figure 4.9: Sentiment polarity of messages.

#### 4.2.2.4 Topic Analysis

Although we here focus on politically related groups, the contents of the messages vary greatly in terms of their topics. Political discussions, product/business marketing, and even humor are some examples. Thus, we further characterized the WhatsApp messages in terms of the topics they convey. To that end, we used Latent Dirichlet Allocation (LDA) [6], a generative statistical model to automatically infer the topics in a collection of documents. We applied LDA to all messages (with misinformation and with unchecked content) jointly, and then compare the distributions of the identified topics in each group of messages, aiming at identifying differences between them. Figure 4.10 presents a flowchart with the steps of our methodology for finding the most frequent topics.

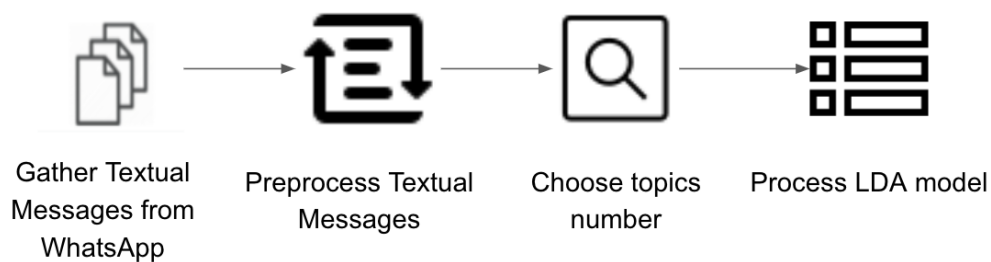


Figure 4.10: Flowchart of Methodology for characterizing WhatsApp textual messages in topics

Specifically, we lowercased and tokenized all the words in the filtered dataset, and removed accents and stopwords using the Portuguese list provided by the Spacy toolkit. We then ran the LDA algorithm using *gensim* [71], a Python library for topic modeling. We chose the best number of topics  $k$  to be returned by the algorithm based on the topic coherence metric [66], which captures whether different topics indeed have few words in common, as is commonly used. Specifically, we ran the LDA algorithm varying the number of topics  $k$  from 2 to 30 and chose the LDA model that produced the highest topic coherence score, which was for  $k = 10$ . These topics are presented in Table 4.5, which shows the most representative words (according to LDA) for each topic. Note that, although our collection methodology does favor political content, we do observe a great variety of topics, characterized by words such as *God*, *life*, *money*, *millions* and *Facebook*.

Table 4.5: Topics inferred by LDA algorithm.

Topic	Most representative words (translated to English)
1	vote, president, Haddad, Lula, Ciro, apply, research, PT, elections, voter
2	no, ant, know, do, person, speak, find, thing, expensive, people
3	say, life, God, do, Lord, day, man, no, good, be
4	country, nation, Brazilian, Brazil, left-wing, political, power, party, govern, right-wing
5	be, laugh, city, governor, senator, yes, state federal, new, big
6	govern, money, do, work, company, millions, year, Brazilian real, pay, receive
7	Bolsonaro, Brazil, say, woman, support, Jair, defend, apply, see, favor
8	be, law, publish, form, education, leave, be, use, project, right, project
9	day, group, Facebook, video, today, folks, chat_whatapp, friend, share, hour
10	year, cop, after, weapon, news, city, arrested, find, crime, where

We then assigned one topic to each message by analyzing the probability of each word in the message belonging to each of the identified topics. We selected the topic with the highest aggregated probability considering all words in the message as its representative topic. Figures 4.11(a) and 4.11(b) show the histograms of topics for the messages with misinformation and for those with unchecked content, respectively.

Clearly, the distribution is much more biased towards fewer topics in the messages with misinformation. The most frequent topic in this group, *Topic 6*, was almost twice as much frequent in the messages with misinformation and is characterized by words such

as *government*, *money*, *do*, and *work*. We found that many messages in this topic labeled as misinformation do indeed carry rumors about government's economic projects in the current or prior term of office. An extract of one such message is: *This is not fake news. It is on the website of the Chamber of Deputies - PT has a project for the confiscation of assets*. It refers to a false project of the political party that had been in Office previously (PT or Work Party), and probably was disseminated aiming at favoring candidates from opposing parties. As this topic is mainly characterized by subjects related to projects, economics and finance, it has no particular political side and their links point to economy news and even false propaganda.

*Topic 1* also has significant presence in the messages with misinformation and is characterized by words such as *Haddad*, *Lula*, and *Ciro* (names of candidates running for president) as well as *vote*, and *president*. Strongly related to the 2018 presidential election, this topic presents information about many candidates, it does not target any particular political side. Those candidates, however were always a target of viral news containing misinformation. The links present in the messages point to news about different candidates and polling surveys results. *Topic 10* however, has words that incite violence and messages that indicate politicians who are supposed to be involved in corruption. Similarly, *Topic 7*, containing mostly words related to *Jair Bolsonaro*, a candidate running for president, was also more frequent in messages with misinformation. This is consistent with reports of how the spread of misinformation in WhatsApp, targeting particular candidates, influenced the 2018 presidential election campaign in Brazil<sup>11</sup>.

One example message related to *Topic 7* is (translated to English): *Bolsonaro proposes mass dismissal of teachers and distance education for all levels*. This fact was learned to be fake afterward, spread with the goal of harming the candidate's campaign. Another example is: *Please listen to what Father Marcelo Rossi talked about the current situation of the country and about Bolsonaro! He gave a class!*. This message refers to a very charismatic and beloved Brazilian priest who allegedly supported candidate *Bolsonaro* in a false audio that went viral. These two messages illustrate how misinformation propagation was used to harm but also to favor the candidate's campaign.

#### 4.2.2.5 Frequent Terms

To further support our analyses of the contents of the WhatsApp messages, Figure 4.12 shows the word clouds of the top 500 most frequent words (translated to English) for

<sup>11</sup><https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html?module=inline>

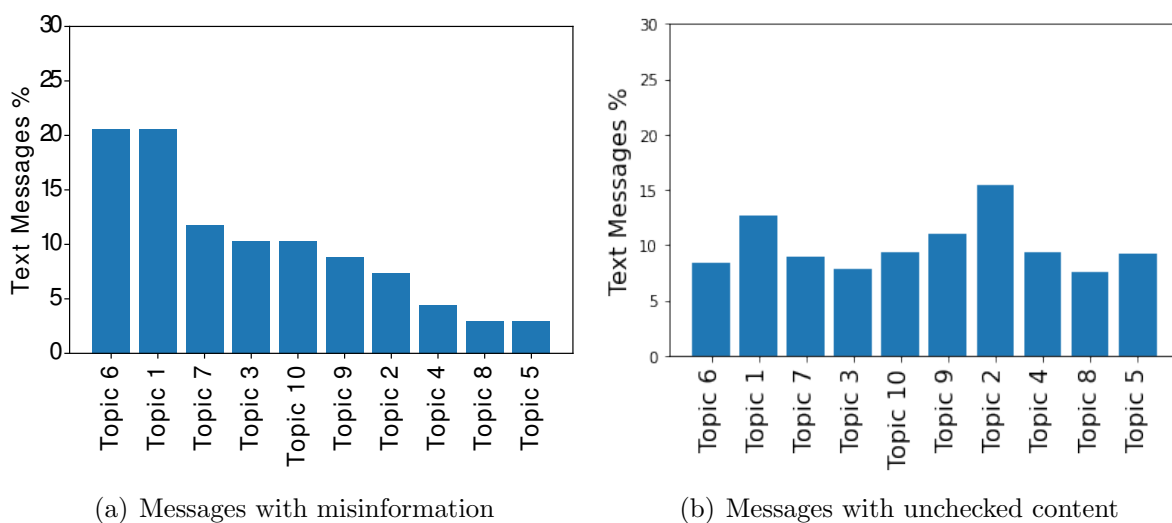


Figure 4.11: Distributions of topics inferred by LDA.

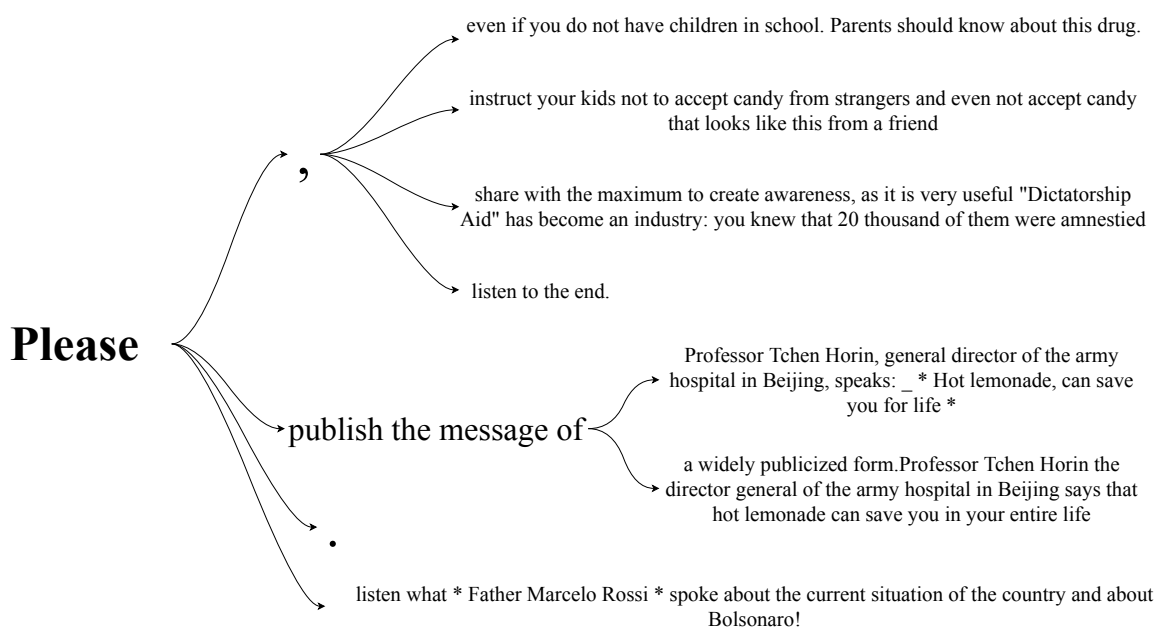
both sets of messages (with misinformation and unchecked content). These word clouds were produced using the Wordle tool<sup>12</sup>. Note the frequent presence of many words related to the topics inferred using the LDA algorithm. Examples are *vote*, *Bolsonaro*, *Brazil*, *Lula* and, *PT*, which are related to the election. Words like *project*, *benefit*, and *income*, clearly related to *Topic 6* (see prior section) are also highlighted in the cloud for messages with misinformation.

We delved further into the contents of messages with misinformation by investigating whether there are particular patterns of word usage (e.g., prefixes or suffixes of sentences) that occur more frequently. Specifically, we used each of the top-50 most frequent words in Figure 4.12(a) as input to the Word tree visualization tool [108]. Given an entry word and a dataset of textual content, this tool generates a tree, with the given entry word as root, showing phrases that branch off from the root across all texts of the dataset.

Figure 4.13 shows one such word tree, rooted by the word *Please* (*Por favor*, in Portuguese). This was the root of the largest number of branches in the set of messages with misinformation. Indeed, as shown in Figure 4.13, we found 7 different phrases starting with the word *Please*. Those phrases were found in messages carrying misinformation, which were shared a total of 33 times in our dataset. We emphasize that, as shown in Figure 4.12(a), these phrases are related to different topics such as a particular candidate (Bolsonaro), health-related issues (e.g., hospital, life), and even a rumor about drugs. This variety of subjects indicates that the use of this particular word *Please* may indeed be a distinguishing feature of misinformation spread in WhatsApp textual messages in general, and not only during the period of elections. Words like *listen*, *publish*, *share*, and *spread* were also found in these phrases. These are words that characterize chain

<sup>12</sup>Available at: <http://www.wordle.net/>



Figure 4.13: Word tree for the word root *Please*.

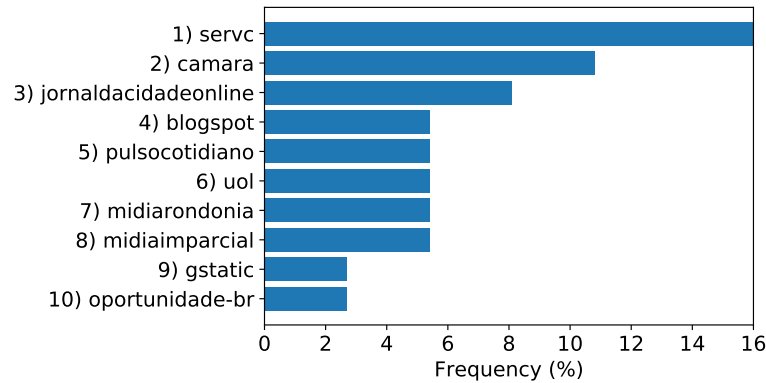
#### 4.2.2.6 Domains Shared

Recall that, as mentioned, a large fraction of messages with misinformation (50% in our dataset) contain URLs to external websites. The same was found in messages with unchecked content, though to a smaller extent (32% of messages). In this section, we analyze the domains these URLs point to. To that end, we first extracted the domains from all identified URLs and then computed the fraction of those URLs pointing to each such domain. Figures 4.14(a) and 4.14(b) show these fractions for the most frequent domains for messages with misinformation and unchecked content, respectively. We note that the most frequent domains shared in messages with misinformation are leading news websites and portals in Brazil (e.g., *globo*, *uol*). Domains with false promotions were also present (e.g., *servc*). Other domains worth mentioning are independent news websites (e.g., *midiaimparcial*) and blogs (*blogspot*).

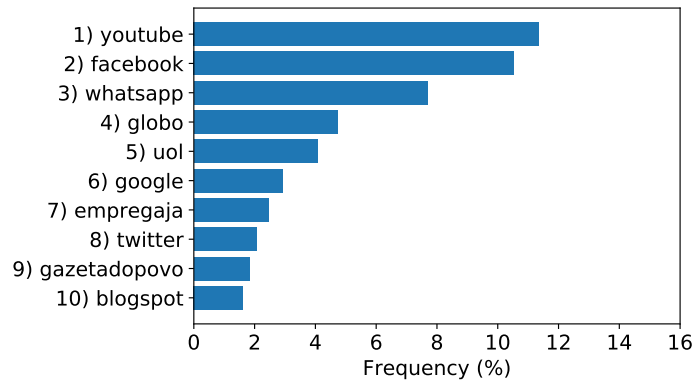
Some of those top domains (i.e. *globo*, *uol*) are leading news websites in Brazil. Those domains were present in the messages with misinformation because they were in messages where people were sharing misinformation checked by those websites. The other domains in Figure 4.14(a) are independent news websites and blogs. Although the percentage of these sites are not high, their presence in a significant part of the messages with misinformation is suspicious. In contrast, social networks (e.g. Facebook, YouTube, Twitter and even WhatsApp itself) are the top domains shared in messages with unchecked content, as shown in Figure 4.14(b). Whatsapp is shared as a invite link



for other WhatsApp groups. Local news websites (e.g., *Gazetadopovo*) and blogs were also present, though with lower frequency.



(a) Messages with misinformation



(b) Messages with unchecked content

Figure 4.14: Most frequent domains in textual messages.

## 4.3 Summary of Results

In this chapter, we have presented the main results of our characterization of the content properties of image and textual messages shared on the monitored groups during the analyzed period (RQ1). In this section we present a summary of the findings, emphasizing differences observed between messages carrying misinformation and messages with unchecked content.

As expected, given the general themes characterizing the monitored groups, we found political content as the major content present in both, images and textual messages. Images with misinformation also have a small percentage of activism and opinion, whereas images with unchecked content have a considerable presence of news and opinion.

Textual messages with misinformation have frequently featured more specific political content, such as government economic projects, candidate rumors, and polling surveys. We also studied psychological features and we found a higher prevalence of attributes like *insight*, *inhibition* and *sexual* on textual messages with misinformation, as also more negative sentiments and a great presence of the word *please*. These patterns suggest a general discourse of calls for action and participation of the users (as in chain messages) (RQ2).

We also checked how textual content and images shared on WhatsApp interplays with the Web. We found that both types of images (with misinformation and the rest) frequently appear in social networks like Twitter (usually on journalistic and candidates' accounts) and Facebook. Images with misinformation, however, also appear in fact-checkers (domains and Twitter accounts), suggesting the occurrence of retractions. In addition, we analyzed domains in URLs present on textual messages and we show a major presence of false promotions, independent news websites, and blogs on messages with misinformation. In textual messages with unchecked content, we found a great presence of social networks and news websites.

## Chapter 5

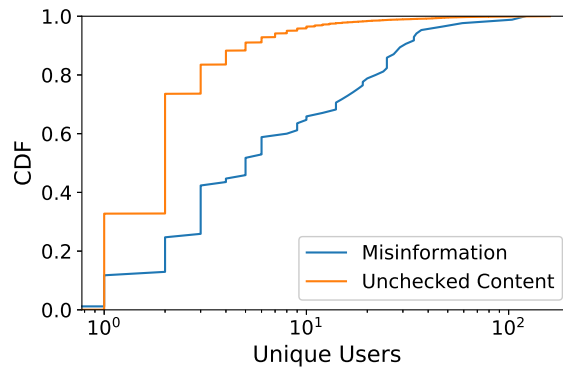
# Propagation Dynamics

In this chapter, we focus on the propagation dynamics of image and textual messages within each group as well as across different groups on WhatsApp (RQ3). We analyze the message reach by quantifying the number of shares (Section 5.1) as well as temporal properties of the spread of a message within WhatsApp (Section 5.2) as well as to and from the Web (Section 5.3). We present results for image and textual messages, separately, identifying differences between messages with misinformation and unchecked content. We finish this chapter with an analysis of the network properties that emerge from the participation of users in different WhatsApp groups (Section 5.4)

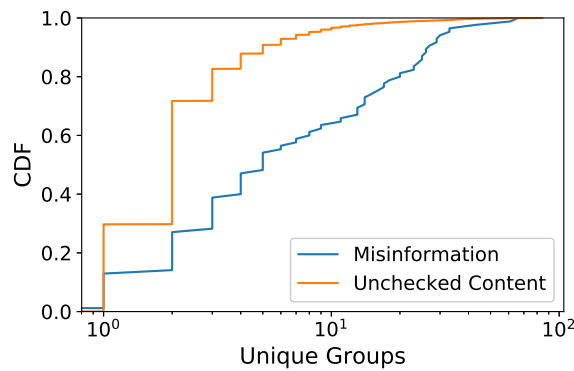
### 5.1 Message Reach

Recall that, as presented in Section 3.3, we do (textual and image) messages with very similar content together and consider them indistinctly duplicates of the same content. In this section, we analyze the reach of each such piece of content by quantifying the number of distinct users who posted the same message, the number of distinct groups in which the same message was posted as well as the total number of copies (shares) of the same message across all analyzed groups. Figures 5.1 and 5.2 show the cumulative distributions of those measures for images and textual messages respectively. Each figure shows two distributions, one for messages with misinformation and the other for messages with unchecked content. For each measure, we compared the two distributions using the Kolmogorov-Smirnov test [59] with 95% confidence level, with the null hypothesis that two samples have the same distribution.

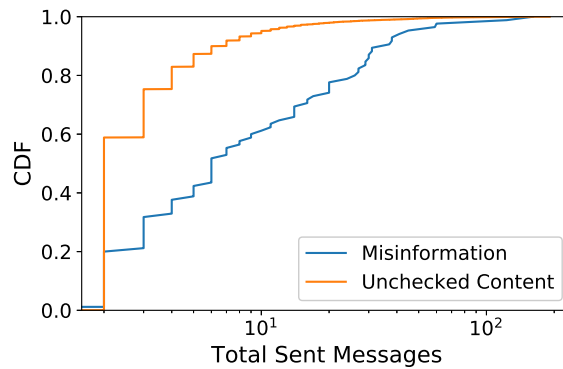
We start by discussing the results for images. As we see in Figure 5.1(a), there is a difference, between the two distributions ( $p$ -value of  $3e-19$ ), roughly 80% of the unchecked images were shared by up to 3 users, although the same fraction of images with misinformation reached a greater number of users (up to 11). We also notice that the images with misinformation tend to reach a much larger number of distinct groups, with a high



(a) Number of unique users per image



(b) Number of unique groups per image



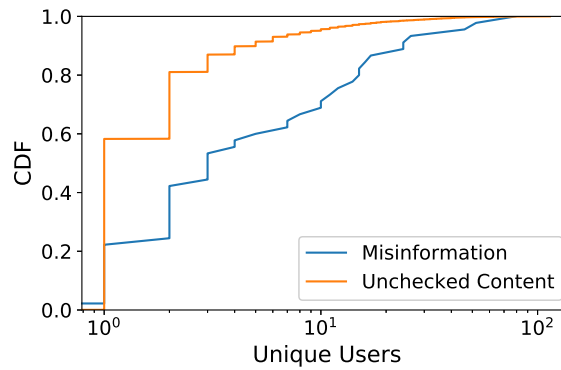
(c) Total number of shares of images

Figure 5.1: Cumulative distributions of the reach of each image in terms of distinct users, distinct groups and total shares.

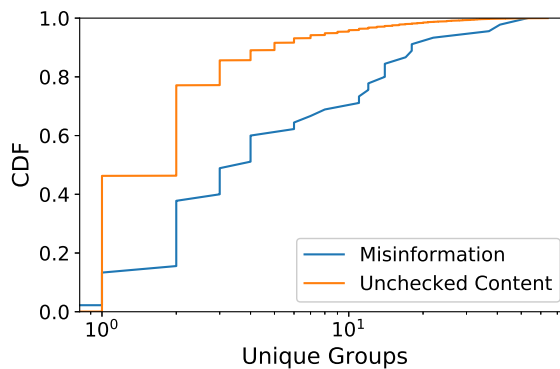
statistical difference ( $p$ -value of  $5e-16$ ) : Figure 5.1(b) indicates that roughly 70% of the images with misinformation were posted in up to 10 groups or, 30% of the images reached more than 10 groups. In contrast, 70% of the images with unchecked content was shared in only up to 2 distinct groups.

As shown in Figure 5.1(c), when we evaluate the total number of shares, we also note a great contrast between images with misinformation and with unchecked content

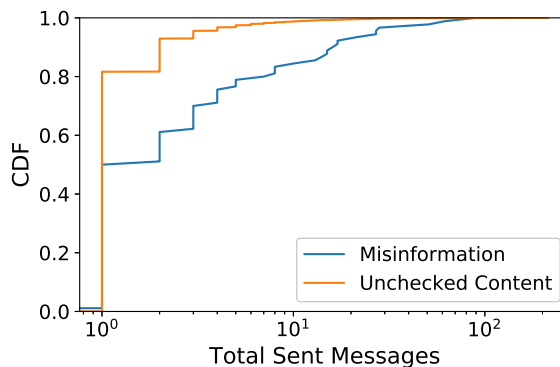
( $p$ -value of  $1e-16$ ). Around 20% of the images with misinformation were shared more than once, and 40% of them were shared more than 10 times. In contrast of this, 80% of the images with unchecked content were shared only 3 times and practically all unchecked content was shared at most 12 times. All these results indicate a much greater spread, reaching more groups and users, on images with misinformation.



(a) Number of unique users per message



(b) Number of unique groups per message



(c) Total number of shares

Figure 5.2: Cumulative distributions of the reach of each textual message in terms of distinct users, distinct groups and total shares.

We now shift our focus to the textual messages. As shown in Figure 5.2(a), roughly

60% of the unchecked messages were shared by up to 2 users, while the same fraction of messages with misinformation reached a much larger number of users (up to 7) showing a difference between the two distributions ( $p$ -value of  $1e-06$ ). Similarly, the messages with misinformation tend to reach a much larger number of distinct groups ( $p$ -value of  $9e-07$ ). Figure 5.2(b) indicates that roughly 80% of the messages were posted in up to 10 groups or, in other way, 20% of the messages reached more than 10 groups. In contrast, 80% of the messages with unchecked content was shared in only up to 2 distinct groups. According to Figure 5.2(c), the distinction between messages with misinformation and with unchecked content is similarly very drastic when it comes to the total number of shares, ( $p$ -value of  $1e-08$ ). Roughly 80% of the latter were shared only once and practically all unchecked content was shared at most 10 times. In contrast, nearly half of the messages with misinformation were shared more than once, and 20% of them were shared more than 10 times. Clearly, images and textual messages with misinformation have a much greater reach in WhatsApp, suggesting a viral behavior within and across the WhatsApp groups.

## 5.2 Propagation Within WhatsApp

In this section, we investigate the spread of a message over time within the monitored WhatsApp groups focusing on messages that were shared at least twice in our dataset. These correspond to 100% and 22% of images with misinformation and unchecked content, respectively, as well as 59% and 16% of textual messages with misinformation and unchecked content. We analyze temporal properties of this spread including message *lifetime* and the time between consecutive shares of the same content, here referred to as *burst time*. Since the computation of burst time disregards the particular group where each share happened, we further analyze the dissemination within and across different groups by analyzing the time interval of a share since the message was first shared in the group (*intra-group time*) and the time interval between the first shares of the same message in different groups (*inter-group time*).

In our analysis, we first identify the set of images and textual messages shared during each monitored period and then compute the aforementioned metrics considering an extended period from April 23<sup>rd</sup> to October 22<sup>nd</sup>, 2018<sup>1</sup>. For each metric, we also compared the two distributions using the Kolmogorov-Smirnov test [59] with 95% confidence level, with the null hypothesis that two samples have the same distribution.

---

<sup>1</sup>The dataset analyzed in this master dissertation is an extract from a larger dataset covering from April to November 2018. We restricted our analysis to the election period only given the greater participation of the group users during that period.

### 5.2.1 Lifetimes

The lifetime of a message is calculated as the time interval between the first and last occurrence of this message in our dataset, thus reflecting how long the message remained being replicated on WhatsApp, *as captured by our dataset*. Figure 5.3 shows the cumulative distributions of the lifetimes (in terms of days) for messages with misinformation and unchecked content and for images and textual messages respectively.

As shown in Figure 5.3(a), we found no statistical difference between the distributions of lifetimes of images with misinformation and images with unchecked content ( $p$ -value of 0.78). For both types of content, around 70% of the images remain in the system for up to 100 hours. However, based on Figure 5.3(b), there is a statistical difference between the distributions of lifetimes of textual messages with misinformation and with unchecked content ( $p$ -value of  $4e-07$ ). Clearly, the textual messages with misinformation tend to remain in the system for much longer: roughly half of the textual messages with misinformation in our dataset had a lifetime of at least 10 days. In contrast, most textual messages with unchecked content remained in the system for up to a single day, and less than 20% of them had lifetimes above 10 days.

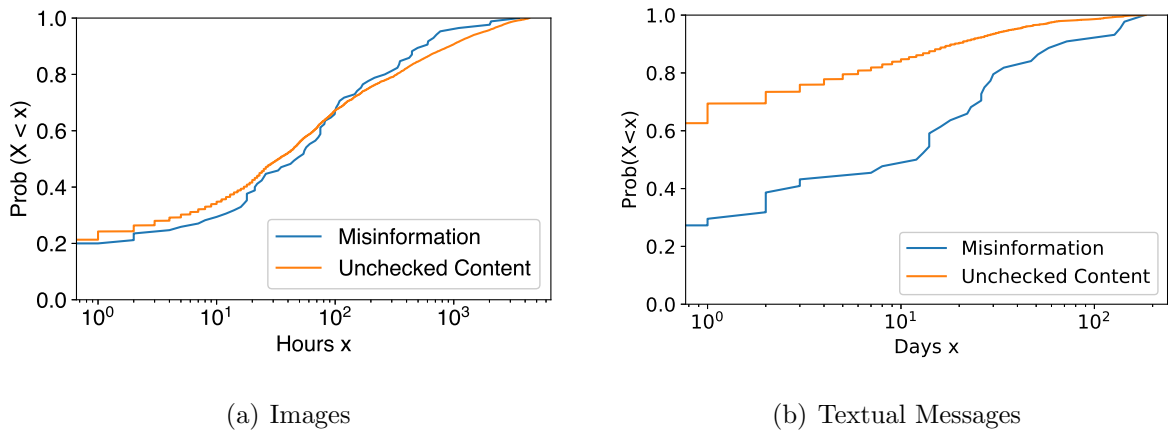


Figure 5.3: Distributions of lifetimes of messages with misinformation and with unchecked content

### 5.2.2 Burst Times

The communication in messenger apps is often extremely fast. Thus, another metric to characterize the temporal dynamics of message propagation is the time interval

between two consecutive shares of the same message (in the same group or in different groups), which we call burst time. Figures 5.4(a) and 5.4(b) show the cumulative distributions of burst time for messages with misinformation and unchecked content for images and textual messages respectively.

Figure 5.4(a) shows that the distributions of burst times are statistically different ( $p$ -value of  $2e-30$ ) for messages with images. Burst times tend to be shorter for images with misinformation, suggesting a faster propagation of this type of content. For example, in 60% of the cases, an image with misinformation is reshared within 100 minutes. The fraction of such burst times reduces to 40% for images with unchecked content.

Additionally, in Figure 5.4(b), we note that the two distributions for textual messages exhibit some distinction in their bodies (smaller values), though differences become unclear for burst times above 100 minutes. That is, for values up to 100 minutes, the burst times tend to be somewhat longer for textual messages with misinformation and messages with unchecked content are reshared faster. For example, around 20% of the messages with unchecked content is reshared within 3 minutes since the last post. In contrast, only 10% of the messages with misinformation are reshared within the same interval. Nevertheless, just like observed for image content, the two distributions are statistically different ( $p$ -value of  $5e-06$ ). We did observe the presence of messages from spammers with promotions and product offers among those with unchecked content. We speculate that those may explain the shorter burst times for such messages, as one may expect that spammers make an effort to publicize their content by resharing it often. Note that, for messages with misinformation, the longer time interval between successive shares of the same message may indeed contribute to the longer lifetimes observed in the previous section.

Once again, we observe differences in the burst time of images and textual messages: while misinformation in images tend to have shorter burst times, the opposite is observed for textual messages. Thus, the misinformation propagation patterns do seem to vary depending on the media type. Extending this analysis to other media types, such as audio and video, is an interesting avenue for future work.

### 5.2.3 Intra and Inter Group Times

We also look into how the same message is disseminated within the same group and across different groups. Our goal is to understand how long it takes for a message to first appear in different groups as well as the time interval since this first appearance and the following shares within the same group. To that end, we define the intra and inter-group



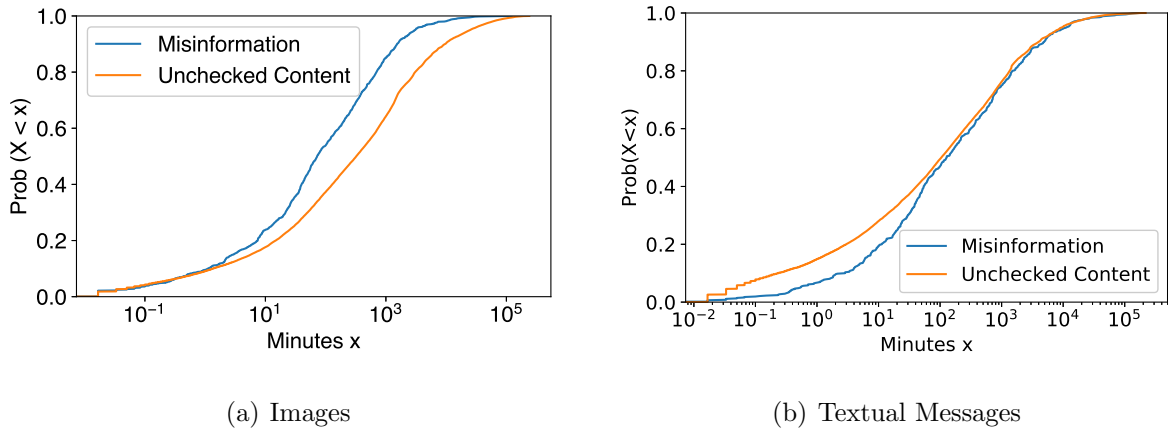


Figure 5.4: Distribution of burst times for messages with misinformation and messages with unchecked content

times. The *intra-group time* is defined as the time interval between the current share of a message and the *first time* the same message was shared in the group. This metric is computed for messages within each group separately, and is restricted to messages shared at least twice in the group. These correspond to 73% and 11% of images with misinformation and unchecked content respectively, as well as for 29% and 9% of textual messages with misinformation and unchecked content. The *inter-group time* is defined as the time interval between the first share of a message in a group and the first share of the message in any group. It captures the time interval between the first appearance of a content in different groups, and is measured only for messages that were shared in at least two groups (87% and 15% for images with misinformation and unchecked content; 48% and 10% for textual messages with misinformation and unchecked content)

We note that the analysis of intra and inter times may help to understand the observed patterns in burst times, since, unlike the latter, the two metrics defined above explicitly capture the structure of groups and its role in the propagation of a message. The cumulative distributions of intra and inter-group times of images are shown in Figures 5.5(a) and 5.5(b), respectively. For both metrics, the distributions for misinformation and unchecked content are statistically different (p-value of  $2e-6$  and  $3e-7$  respectively), though the differences are not very expressive. In general, we observe that images with unchecked content tend to spread somewhat faster within the groups (shorter intra-group times) but tend to take longer to cross the boundaries between groups (longer inter-group times). For example, as shown in 5.5(a) 60% of the images with misinformation have intra-group times of up to 100 hours, whereas for unchecked content the threshold is 50 hours. In contrast, the inter-group times of 60% of unchecked images is 70 hours, whereas for images with misinformation the corresponding value is only 50 hours (see Figure 5.5(b)).

The cumulative distributions of intra and inter-group times of textual messages are shown in Figures 5.6(a) and 5.6(b), respectively. As the figures show, the patterns

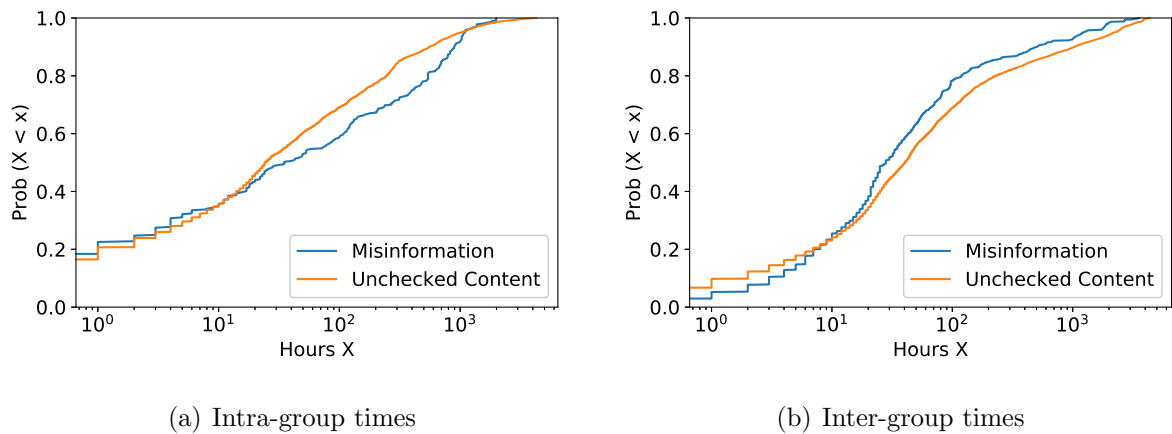


Figure 5.5: Distribution of inter and intra-group for images with misinformation and messages with unchecked content.

are completely the opposite of what was observed for images. Within each group, the shares of messages with misinformation tend to be somewhat more concentrated in time, happening faster. As Figure 5.6(a) shows, in approximately 50% of the cases, a message with misinformation is reshared within 10 hours since the first time it appeared in the group. For messages with unchecked content, this fraction is smaller than 40%. We also note a statistical difference between the distributions of intra-group time of textual messages with misinformation and with unchecked content (p-value of 0.01). In contrast, Figure 5.6(b) shows that crossing the group boundaries takes longer for messages with misinformation: in only 20% of the cases, they reappear in a different group within 10 hours. However, for messages with unchecked content, 30% of the inter-group times are within the same limit. Those distributions also present a statistical difference (p-value of  $1e-16$ ).

By contrasting these results with those reported in the previous section, we conclude that although the overall spread of textual messages with misinformation is somewhat slower (greater burst times), images, in general, spread slower within particular groups taking a faster time to propagate across different groups. Textual messages, however, spread faster within particular groups, taking longer to propagate across different groups.

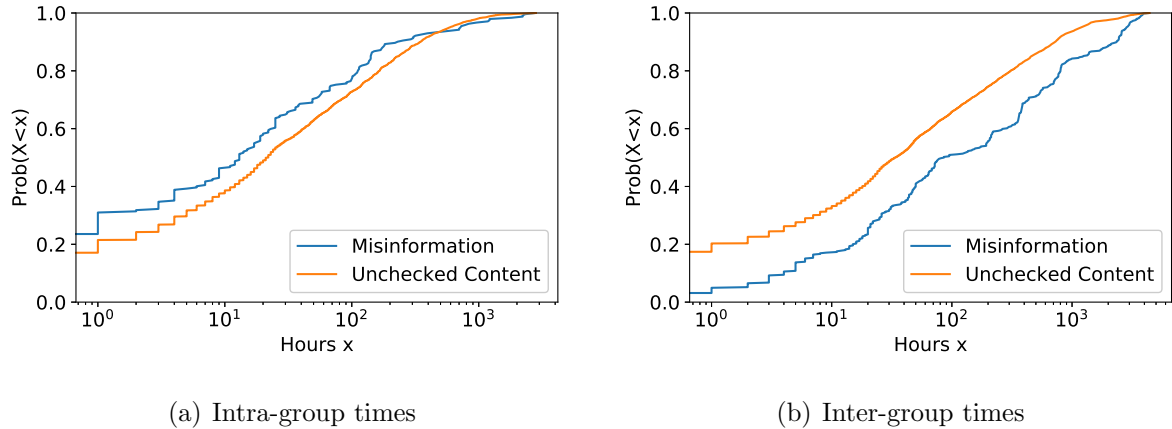


Figure 5.6: Distribution of inter and intra-group for textual messages with misinformation and messages with unchecked content.

### 5.3 Propagation to and from the Web

We also analyze the propagation of images across the boundaries between WhatsApp and the Web. Specifically, we analyze the difference between the time an image was first shared on a monitored group and the time when it was indexed by Google. The latter is taken as an estimate of the time it first appeared on the Web. A positive difference suggests that the image was first shared in one of the monitored groups and then published on the Web. A negative difference may suggest the image was first posted on the Web<sup>2</sup>.

Figure 5.7 shows the cumulative distribution of such time differences for images with unchecked content and images with misinformation. We compared the two distributions using the Kolmogorov- Smirnov test [59] with 95% confidence level, with the null hypothesis that the two samples have the same distribution. We found that the distributions are clearly different ( $p$ -value of  $2.4e-47$ ). The vast majority (95%) of images with unchecked content were first posted on the Web (negative intervals). Only 3% of them were shared first on the monitored groups (positive intervals) whereas 2% appeared on both Web and WhatsApp on the same day. In contrast, only 45% of the images with misinformation were shared first on the Web, 20% of them were shared on both platforms on the same day, and 35% were shared first on the WhatsApp group. These results seem to suggest that WhatsApp acted as a source of images with misinformation during the election campaign period. This observation is in alignment with the *Trumpet of Amplification* argument that states that WhatsApp groups as other closed or semi-closed

<sup>2</sup>Our analysis is constrained by the view of WhatsApp provided by our dataset. We restrict this analysis only to images as identifying occurrences of the same textual content elsewhere on the Web would be significantly harder.

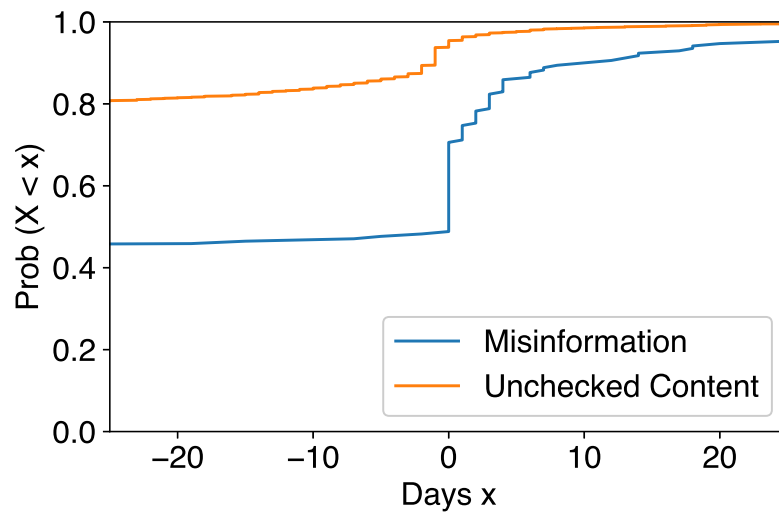
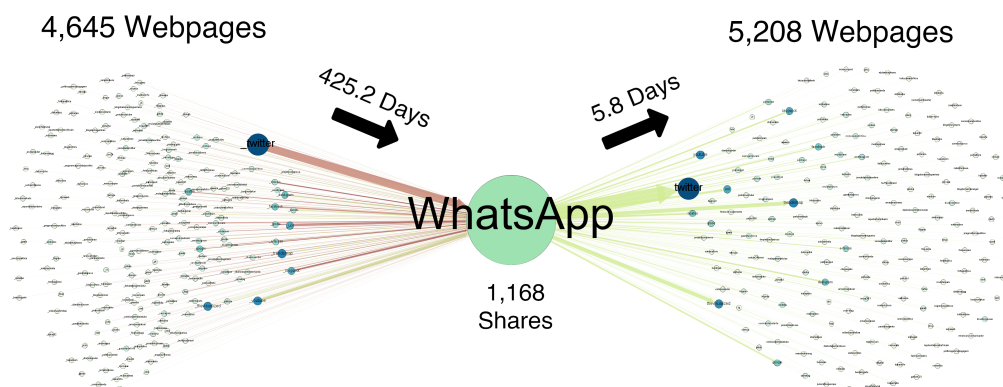


Figure 5.7: Cumulative Distribution Function for Temporal Propagation on Web of WhatsApp Images

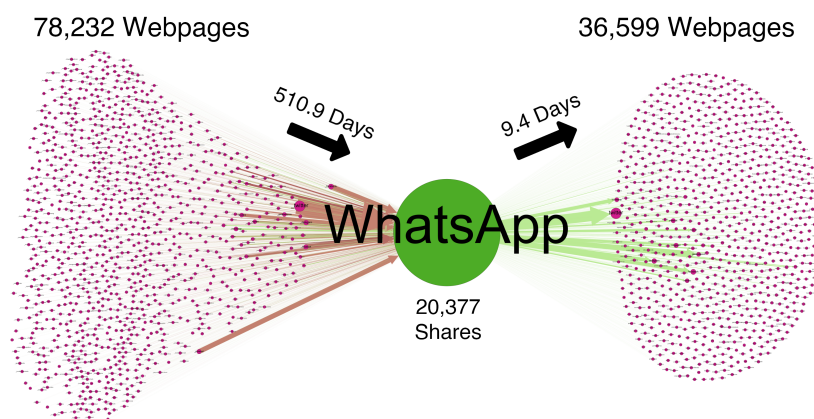
communities (4chan, Facebook groups, etc) often act as sources for the dissemination of misinformation [106].

To further investigate the sharing of image content on the monitored WhatsApp groups and on the Web, we propose a visualization by means of a directed network, as shown in Figure 5.8. The network contains a central node representing WhatsApp (i.e., the monitored groups); the other nodes represent Web domains in which the images shared on WhatsApp also appeared. A directed edge from a node/domain to the central node implies that an image first appeared on that domain and later it was shared on WhatsApp. A directed edge from the central node to a node/domain implies the opposite. To improve readability, we plot nodes representing domains in which the images appeared before being shared on WhatsApp on the left of the central node, and nodes representing domains in which the images appeared after being shared on WhatsApp on its right. The size of a node representing a domain captures the number of webpages in that domain in which images shared on WhatsApp appeared. The color of an edge represents the average time difference between the first appearance of an image on WhatsApp and on the specific domain, considering all images posted on that domain (green is faster than red). We emphasize that this representation captures the temporal ordering of the first appearance of an image within WhatsApp and on the Web, as captured by our dataset. Although it may provide hints about the propagation of image across the boundaries between WhatsApp and the Web, we cannot claim they map exactly the actual information flow.

Figure 5.8 shows the network representations for images with misinformation and images with unchecked content. In addition to the network itself, each figure shows, for each group of domains, the total numbers of pages containing shared images as well as



(a) Images with misinformation.



(b) Images with unchecked content.

Figure 5.8: Network representation of images shared on WhatsApp and on the Web.

the average time interval between the first appearance of an image on the Web and on WhatsApp. We note that images that were first published on the Web take much longer to reach the WhatsApp groups (more than a year) than the other way around (only a few days) for both sets of images. The average time interval is 73 times longer for images with misinformation and 54 times longer for images with unchecked content. Also, in general, images with misinformation cross the boundaries between WhatsApp and the Web much more quickly: 425 days from the Web to WhatsApp and less than 6 days from WhatsApp to the Web, on average (as opposed to 511 and 9 days, respectively, for images with unchecked content). Moreover, the numbers of domains (and webpages) on both sides of the central node are much more balanced for images with misinformation. This suggests, once again, that images with misinformation are much more often spread from the WhatsApp groups to the rest of the Web than images with unchecked content. We note, however that fact checkers sites are included in those webpages and no domain were excluded for this analysis.

## 5.4 Network Structures

In this section, we analyze sharing patterns on the selected WhatsApp groups by studying the structure of the networks that emerge from the participation of users in different groups (Section 5.4.1). In this analysis, we focus only on the sharing of *images*, but we expect similar patterns to emerge for textual messages as well. We also study the network that emerge from the sharing of misinformation (Section 5.4.2).

### 5.4.1 General Network Properties

We modeled the interactions across groups by means of two network models, one at the group level and one at the user level. That is, we built a *group network* where each node represents one monitored group and edges are added connecting groups that have at least one member in common sharing image content. Figure 5.9 shows the group networks built. The size of each node represents the number of users who shared the same content in the group. Although many groups are somewhat isolated or weakly connected to the rest, we do note the presence of several clusters of groups which are strongly interconnected by sharing many members in common. This might facilitate the flow of information across group boundaries.

We also modeled the relationship between users by building a *user network* where each node is a user and an edge is added between two nodes if the corresponding users have shared image content in at least one group in common. Node size represents the number of groups in which the user shared images. The user networks are naturally larger and harder to visualize. For illustration purposes, Figure 5.10 shows a subgraph of the network built, with 5,700 nodes. The network structure of the groups is evidenced by the clusters formed. We note a large number of users blending together connecting to each other inside those groups. Most users indeed form a single cluster, connecting mostly to other members of the same community. On the other hand, there are also a few users who serve as bridges between two or more groups linked by multiple users at the same time. Furthermore, a few users work as big central hubs, connecting multiple groups simultaneously. Lastly, some groups have a lot of users in common, causing these groups to be strongly inter-connected, making it even difficult to distinguish them.

To better understand the properties of these graphs, Table 5.1 shows various network metrics computed for the group and user networks. It presents numbers of nodes

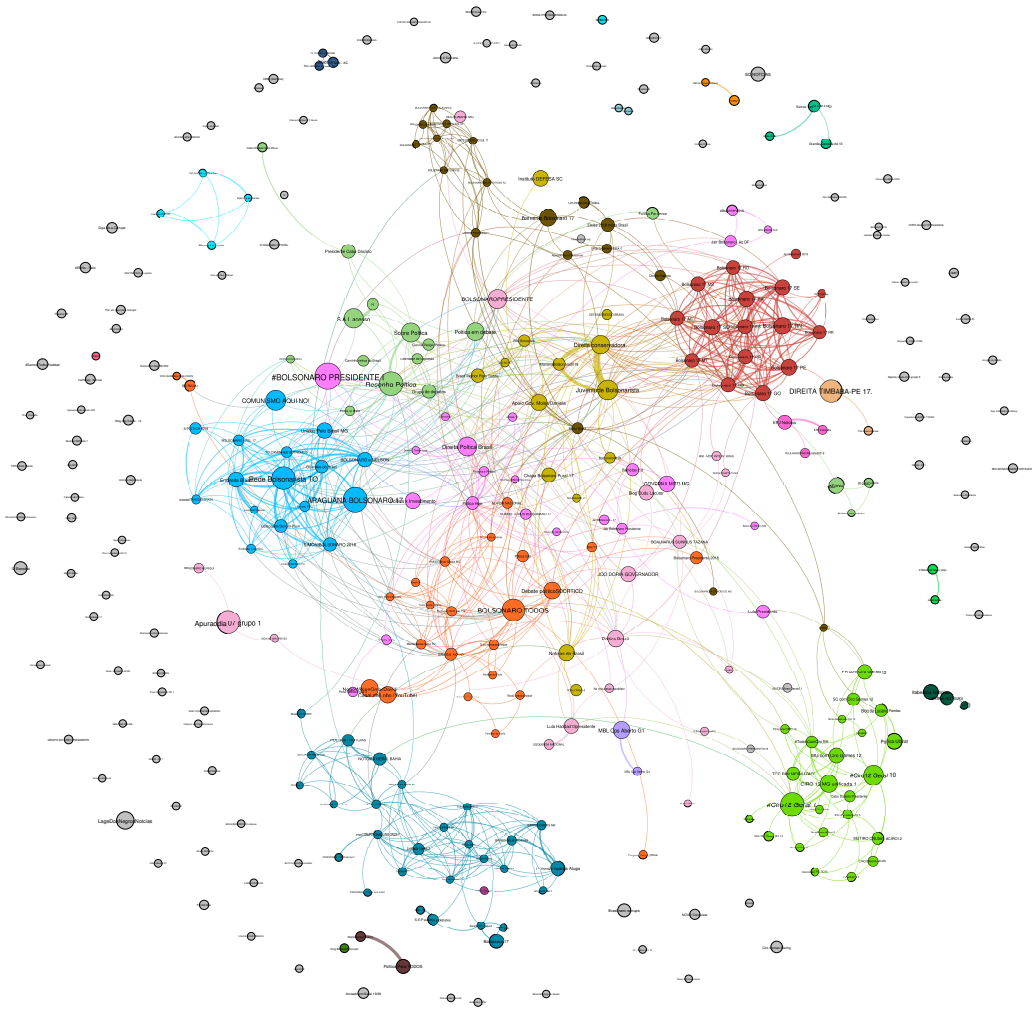


Figure 5.9: WhatsApp Network of Users with public political groups in common

and edges, average node degree, network diameter, average path length (APL), network density, and the size of the largest connected component (LCC).

Table 5.1: Network metrics for WhatsApp graphs.

	#Nodes	#Edges	Avg. Degree	Diameter	APL*	Density	LCC**
Group Network	333	842	5.057	8	3.459	0.015	206
User Network	10,860	492,217	90.91	9	3.952	0.008	8,934

\*Average Path Length. \*\*Largest Connected Component.

We note that the group network is complex and densely connected, with clusters (i.e., communities) of groups and edges emerging between them, and a large fraction of nodes belonging to the largest connected component (62%). Despite such differences, the average path length between the groups is 3.46. The network density (ratio of the number of edges in the graph to the maximum number of edges possible) is low (under 2%) We observe similar properties in the user network with a small average shortest path length (3.95) and higher largest connected component (82% of the users).

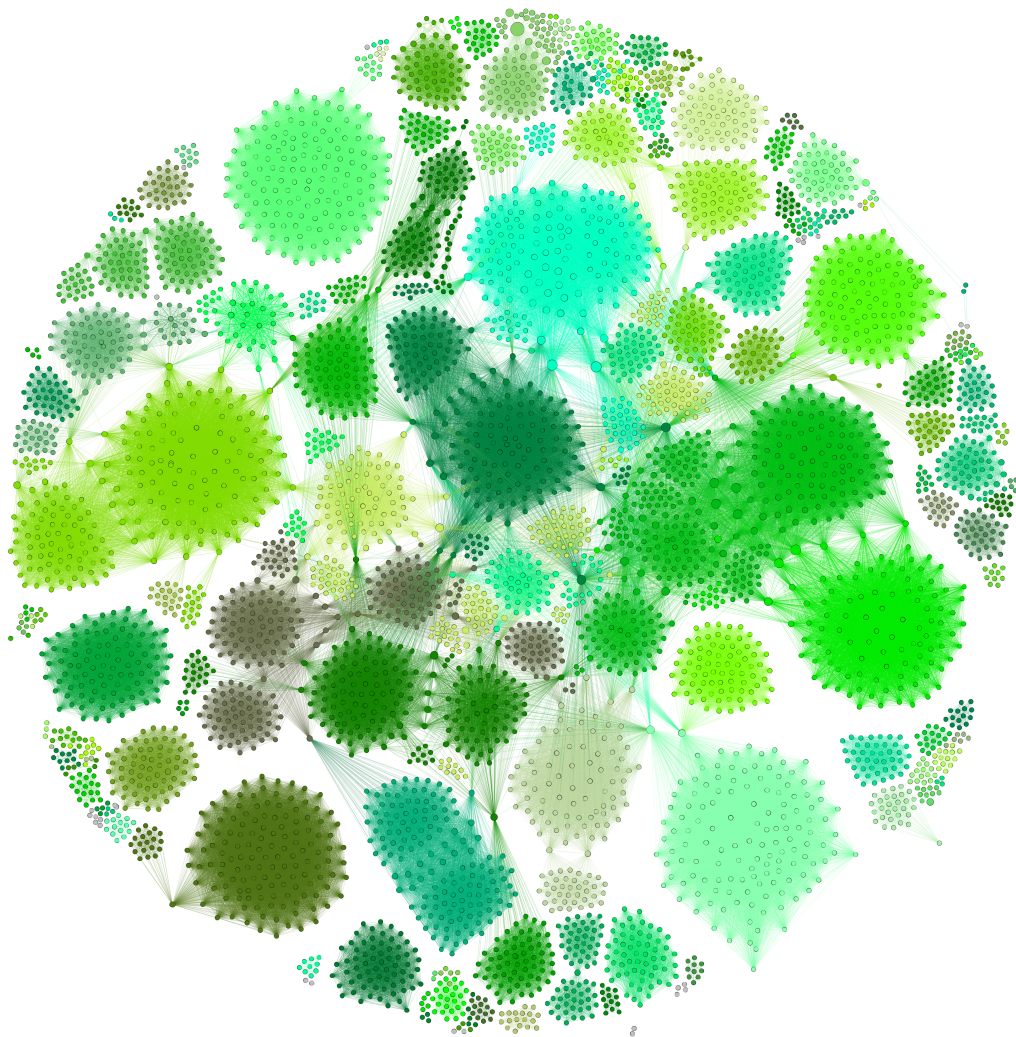


Figure 5.10: WhatsApp Network of Users with public political groups in common

Therefore, as illustrated in Figures 5.9 and 5.10 and in Table 5.1, WhatsApp is more than just a mobile network that provides end-to-end encrypted communication between two users. It exhibits network properties as Caveman Model [109], very similar to many other social networks such as Twitter or Facebook, connecting thousands of users and having the potential to make a piece of information become viral. We highlight, however, that those networks presented are incomplete in terms of edges, once we only considered posts of images.



### 5.4.2 Misinformation Network

We analyzed the propagation of images with misinformation on the WhatsApp groups by building a network model representing the groups in which the messages with misinformation *first appeared*. Specifically, we built a directed graph where each node represents a group and a directed edge from node  $A$  to node  $B$  was added if the same image with misinformation was first shared in group  $A$  and then appeared in group  $B$ . To build this graph we considered only groups in which at least 2 distinct images with misinformation were shared during the period. The weight of an edge is defined as the number of images containing misinformation that were first shared in a group and then co-occurred in the other. The size of a node represents the number of images with misinformation posted on that particular group while the color represents the sum of the outgoing edges, that is, the total number of images that were “first seen” in that group and then spread to the rest of the network.

Figure 5.11 shows the network of propagation of images with misinformation in the monitored WhatsApp groups during the analyzed period. Note that some nodes are darker (larger out-degree) than others, suggesting they are the main “seeds” of the images with misinformation in the graph. It is worth noting that the group in which the largest number of images with misinformation first appeared (largest node) is indeed the group with the largest number of users and the largest number of images shared in general. Yet, we note that some large nodes have very light colors (e.g. “*ARAGUANA BOLSONARO 1*” and “*BOLSONAROPRESIDENTE*”), meaning that although many images containing misinformation were shared in them, they acted more as receptors than seeds, since their out-degrees are small. These results seem to suggest that fewer groups are responsible for the spreading of a large fraction of the images with misinformation in WhatsApp.

## 5.5 Summary of Results

In this chapter, we have presented the main results of our analyses of the propagation dynamics of image and textual messages within the monitored groups (as well as to and from the Web) during the analyzed period (RQ3). In this section we present a summary of the findings, emphasizing differences observed between messages carrying misinformation and messages with unchecked content.

We found that for both images and textual content, the messages with misinfor-



---

measured that messages with misinformation cross boundaries to/from the Web faster than messages with misinformation. Finally, we analyzed the network of misinformation and we presented a graph in Figure 5.11 showing how misinformation spreads between the groups we monitored. The results suggest the presence of a few nodes acting as potential "seeds" of messages with misinformation, while the other nodes act more as receptors of misinformation.

## Chapter 6

# Conclusions and Future Work

This master thesis presents an analysis of messages shared on publicly accessible WhatsApp groups related to politics during the first round of the 2018 Brazilian general elections campaign in Brazil. We have analyzed the properties and propagation dynamics of messages disseminated in a number of politically oriented WhatsApp groups. Our study was driven by the goal of identifying properties that distinguishing messages containing previously reported misinformation from the rest, for both images and textual messages. To that end, we relied on a dataset of fake news reported by six Brazilian fact-checking websites, identifying their presence in the WhatsApp messages analyzed (images and textual messages).

We found that images are the most popular type of media content shared on this platform during the period of our analysis. We analyzed the main features present in images with misinformation contrasting it with the patterns observed for the other (unchecked) image contents. Moreover, by manually labeling these images, we found the frequent presence of activism, and personal opinions and much of this content came from other social networks, independent websites, and blogs on images with misinformation. We characterized the propagation dynamics of these images and we found that images with misinformation tend to be reshared within shorter time intervals, spreading faster when crossing boundaries in distinct groups but taking longer when reshared in the same group. We also offered insights into how information may propagate to/from the Web and reported that images with misinformation are often shared first on WhatsApp and then on the Web. This observation suggests that WhatsApp may have been a relevant source of images with misinformation to the Web during the analyzed period.

We also analyzed textual messages and our results revealed a number of interesting findings. With respect to textual properties, we found only small differences in message sizes as messages with misinformation tend to be slightly smaller (especially in number of words). This may be partially due to the larger presence of URLs in their contents. By performing topic modeling, we also identified that textual messages with misinformation are more concentrated on fewer topics, related to presidential candidates and government projects. The prevalence of such topics was confirmed by a higher frequency of words related to them, and this is consistent with the general theme of the monitored group as

---

well as the analyzed period. Moreover, the analysis of the psychological elements indicated a frequent presence of the cognitive process of *insight* in the messages with misinformation. This attribute is characterized by words such as *attention*, *warning*, *look*, *listen* which are often used in chain messages. We also noted the frequent presence of phrases starting with the word *Please*, used in relation to various subjects, which may also be a feature of chain messages. Finally, despite the differences being small, we do find that the contents of messages with misinformation tend to be more negative, in agreement with previous analysis of misinformation [115].

Our analyses of the propagation dynamics revealed a much more viral spread of misinformation content, as such messages are shared more times, by a larger number of users and in more groups for both images and textual messages, consistently with our results for images. However, we did observe some differences across the different media types. Textual messages with misinformation tend to spread faster within particular groups, but take longer to propagate across different groups, which results in such messages lasting longer on WhatsApp. These results are in contrast with our study of the same time propagation metrics of misinformation in images, suggesting that the propagation dynamics of misinformation may indeed depend on the type of media used to convey the information.

As a complement, we characterized the network structure of the monitored WhatsApp groups, showing how they connect with each other and offering insights into how information may propagate between. Our results suggest that, even though WhatsApp was not designed to be a social network, the networks that emerge from user participation and content sharing do have properties similar to previously analyzed social networks (e.g., short diameter, large connected component, low network density and small average path length) [23, 52]. We also found that there are a few nodes (groups) where misinformation is sent first, suggesting they can be the main seeds of misinformation in the network of groups from our dataset.

We emphasize that although our findings were observed on a particular dataset and thus might be influenced by its collection methodology (e.g., focus on political groups, the particular time period monitored), they might generalize, to some extent, to other WhatsApp groups and periods. For example, although the observed particular topics are biased by our collection methodology, the concentration of misinformation on fewer (more catchy and controversial) topics may be expected in general, and so are the longer lifetimes.

This study offers a first step towards understanding how misinformation disseminates in textual content and images, the two major message types on WhatsApp. We hope it motivates follow-up efforts covering other datasets, time periods, WhatsApp groups and media types. For example, characterizing the spread of audios and videos, particularly those carrying misinformation, may be an interesting avenue to pursue, given the increas-

ing use of this feature on WhatsApp. Exploring the analyzed features in the design of automatic mechanisms for detecting misinformation on WhatsApp is also a promising future work.

More broadly, we expect this study to drive follow-up investigations covering other types of content as well as delving further into the interplay between WhatsApp groups and the Web as channels for information propagation.

# Bibliography

- [1] Khalid AJ Al Khaja, Alwaleed K AlKhaja, and Reginald P Sequeira. Drug information, misinformation, and disinformation on social media: a content analysis study. *Journal of public health policy*, 39(3):343–357, 2018.
- [2] Jonathan Albright. The #election2016 micro-propaganda machine. <https://medium.com/@d1gi/the-election2016-micro-propaganda-machine-383449cc1fba>, 2016.
- [3] Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabrício Benvenuto, Krishna P Gummadi, and Adrian Weller. Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 10–16. ACM, 2018.
- [4] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.
- [5] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, 21(11), 2016.
- [6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [7] Dan Bouhnik and Mor Deshen. Whatsapp goes to school: Mobile instant messaging between teachers and students. *Journal of Information Technology Education: Research*, 13(1):217–231, 2014.
- [8] Victor S Bursztyn and Larry Birnbaum. Thousands of small, constant rallies: A largescale analysis of partisan whatsapp groups. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019.
- [9] Josemar Alves Caetano, Jaqueline Faria de Oliveira, Hélder Seixas Lima, Humberto T Marques-Neto, Gabriel Magno, Wagner Meira Jr, and Virgílio AF Almeida. Analyzing and characterizing political discussions in whatsapp public groups. *arXiv preprint arXiv:1804.00397*, 2018.

- [10] Josemar Alves Caetano, Gabriel Magno, Evandro Cunha, Wagner Meira Jr, Humberto T Marques-Neto, and Virgilio Almeida. Characterizing the public perception of whatsapp through the lens of media. In *Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media (RDSM 2018)*, 2018.
- [11] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T. Marques-Neto, and Virgilio Almeida. Characterizing attention cascades in whatsapp groups. In *Proceedings of the 10th ACM Conference on Web Science, WebSci '19*, pages 27–36, New York, NY, USA, 2019. ACM.
- [12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.
- [13] J. Constine. Whatsapp hits 1.5 billions monthly users. <https://techcrunch.com/2018/01/31/whatsapp-hits-1-5-billion-monthly-users-19b-not-so-bad/>, 2018.
- [14] Evandro Cunha, Gabriel Magno, Josemar Caetano, Douglas Teixeira, and Virgilio Almeida. Fake news as we feel it: perception and conceptualization of the term “fake news” in the media. In *International Conference on Social Informatics*, pages 151–166. Springer, 2018.
- [15] Erond L Damanik. Middle class, whatsapp, and political orientation: The election of north sumatera governor, 2018. In *1st International Conference on Social Sciences and Interdisciplinary Studies (ICSSIS 2018)*. Atlantis Press, 2019.
- [16] S. Darlington. As ‘Lula’ Sits in Brazil Jail, Party Nominates Him for President. <https://www.nytimes.com/2018/08/05/world/americas/lula-brazil-election-luiz-inacio-lula-da-silva.html>, 2018.
- [17] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, Kasim Selcuk Candan, Lexing Xie, and Aisling Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.
- [18] Philippe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro OS Vaz de Melo, and Fabrício Benevenuto. Can whatsapp counter misinformation by limiting message forwarding? pages 372–384, 2019.
- [19] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. Why rumors spread fast in social networks. *Communications of the ACM*, 55(6):70–75, 2012.



- [20] Financial Express. Whatsapp now has 1.5 billion monthly active users, 200 million users in india. <https://www.financialexpress.com/industry/technology/whatsapp-now-has-1-5-billion-monthly-active-users-200-million-users-in-india/1044468/>, 2018.
- [21] Gowhar Farooq. Politics of fake news: How whatsapp became a potent propaganda tool in india. *Media Watch*, 9(1):106–117, 2017.
- [22] Cristobal Fernandez-Robin, Diego Yañez, and Scott McCoy. Intention to use whatsapp. In *Artificial Intelligence*. IntechOpen, 2019.
- [23] Emilio Ferrara. A large-scale community structure analysis in facebook. *EPJ Data Science*, 1(1):9, 2012.
- [24] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8), 2017.
- [25] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, June 2016.
- [26] J. Fleiss, B. Levin, and M. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.
- [27] Adam Fourney, Miklos Z Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. Geographic and temporal trends in fake news consumption during the 2016 us presidential election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2071–2074. ACM, 2017.
- [28] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- [29] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. *WOSN*, 10:3–11, 2010.
- [30] Kiran Garimella and Gareth Tyson. Whatapp doc? a first look at whatsapp public group data. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.
- [31] Tali Gazit and Noa Aharony. Factors explaining participation in whatsapp groups: an exploratory study. *Aslib Journal of Information Management*, 70(4):390–413, 2018.
- [32] Vindu Goel, Suhasini Raj, and Priyadarshini Ravichandran. How whatsapp leads mobs to murder in india.

- <https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html>, 2018.
- [33] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B Everett, et al. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–21. ACM, 2018.
- [34] Juliana Gagnani. Pesquisa inédita identifica grupos de família como principal vetor de notícias falsas no whatsapp. <https://www.bbc.com/portuguese/brasil-43797257>, 2018.
- [35] 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.
- [36] Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 1145–1152. ACM, 2012.
- [37] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.
- [38] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.
- [39] Peter Herson. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139, 1995.
- [40] Thi Bich Ngoc Hoang and Josiane Mothe. Predicting information diffusion on twitter—analysis of predictive features. *Journal of computational science*, 28:257–264, 2018.
- [41] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.
- [42] Jun Ito, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama. Assessment of tweet credibility with lda features. In *Proceedings of the 24th International Conference on World Wide Web*, pages 953–958. ACM, 2015.

- [43] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2972–2978. AAAI Press, 2016.
- [44] Anna-Katharina Jung, Milad Mirbabaie, Björn Ross, Stefan Stieglitz, Christoph Neuberger, and Sanja Kapidzic. Information diffusion between twitter and online media. 2018.
- [45] Pooja Khurana and Deepak Kumar. Sir model for fake news spreading through whatsapp. In *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIOTCT)*, pages 26–27, 2018.
- [46] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM, 2018.
- [47] Young Mie Kim, J. Hsu, D. Neiman, C. Kou, L. Bankston, S. Kim, R. Heinrich, R. Baragwanath, and G. Raskutti. The stealth media? groups and targets behind divisive issue campaigns on facebook. *Political Communication*, 0(0):1–27, 2018.
- [48] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
- [49] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.
- [50] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.
- [51] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.
- [52] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
- [53] D Lazer, M Baum, N Grinberg, L Friedland, K Joseph, W Hobbs, and C Mattsson. Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy*, 2, 2017.

- [54] Paula Leite. In brazil, whatsapp is the main carrier for fake news; in the us, it's facebook. <https://www1.folha.uol.com.br/internacional/en/opinion/2018/10/in-brazil-whatsapp-is-the-main-carrier-for-fake-news-in-the-us-its-facebook.shtml>, 2018.
- [55] Mallory Locklear. Researchers say facebook's anti-fake news efforts might be working. <https://www.engadget.com/2018/09/14/facebook-fake-news-efforts-working/>, 2018.
- [56] Tessa Lyons. Hard questions: How is facebook's fact-checking program working? <https://newsroom.fb.com/news/2018/06/hard-questions-fact-checking>, 2018.
- [57] Matheus Magenta, Juliana Gragnani, and Felipe Souza. How whatsapp is being abused in brazil's elections, 2018.
- [58] Anang Marfianto and Imam Riadi. Whatsapp messenger forensic analysis based on android using text mining method. *Int. J. Cyber-Security Digit. Forensics*, 7(3):319–327, 2018.
- [59] F. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
- [60] Kehinde Funmilayo Mefolere. Whatsapp and information sharing: prospect and challenges. *International Journal of Social Science and Humanities Research*, 4(1):615–625, 2016.
- [61] Philippe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. Whatsapp monitor: A fact-checking system for whatsapp. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 676–677, 2019.
- [62] Philippe F Melo, Daniel H Dalip, Manoel M Junior, Marcos A Gonçalves, and Fabrício Benevenuto. 10sent: A stable sentiment analysis method based on the combination of off-the-shelf approaches. *Journal of the Association for Information Science and Technology*, 2019.
- [63] Johnnatan Messias, Lucas Schmidt, Ricardo Rabelo, and Fabrício Benevenuto. You followed my bot! transforming robots into influential users in twitter. *First Monday*, 18(7), July 2013.
- [64] V. Monga and B. Evans. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE Transactions on Image Processing*, 15(11):3452–3465, 2006.

- [65] Ashish Moon and T Raju. A survey on document clustering with similarity measures. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(11):599–601, 2013.
- [66] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.
- [67] D. Phillips. Jair bolsonaro: Brazil presidential frontrunner stabbed at campaign rally. <https://www.theguardian.com/world/2018/sep/06/brazil-jair-bolsonaro-far-right-presidential-candidate-stabbed>, 2018.
- [68] Peter Pomerantsev and Michael Weiss. *The menace of unreality: How the Kremlin weaponizes information, culture and money*. Institute of Modern Russia New York, 2014.
- [69] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 155–158. International World Wide Web Conferences Steering Committee, 2018.
- [70] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and tracking political abuse in social media. *ICWSM*, 11:297–304, 2011.
- [71] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proc. of the Workshop on New Challenges for NLP Frameworks (LREC)*, 2010.
- [72] Julio Reis, Fabricio Benevenuto, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. Breaking the news: First impressions matter on online news. In *Proc. of the Int’l AAAI Conference on Web and Social Media (ICWSM)*, 2015.
- [73] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2), 2019.
- [74] 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 *Proceedings of the 10th ACM Conference on Web Science, WebSci ’19*, pages 225–234, New York, NY, USA, 2019. ACM.

- [75] 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 World Wide Web Conference*, WWW '19, pages 818–828, New York, NY, USA, 2019. ACM.
- [76] Gustavo Resende, Johnnatan Messias, Márcio Silva, Jussara Almeida, Marisa Vasconcelos, and Fabrício Benevenuto. A system for monitoring public political groups in whatsapp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, pages 387–390, New York, NY, USA, 2018. ACM.
- [77] Filipe N Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P Gummadi, and Elissa M Redmiles. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 140–149. ACM, 2019.
- [78] Manoel Horta Ribeiro, Pedro H. Calais Guerra, Wagner Meira Jr., and Virgilio Almeida. "everything i disagree with is #fakenews": Correlating political polarization and spread of misinformation. In *Data Science + Journalism Workshop KDD 2017*, Halifax, Canada, 2017.
- [79] Marian-Andrei Rizoiu, Timothy Graham, Rui Shang, Yifei Zhang, Robert Ackland, Lexing Xie, et al. # debatenight: The role and influence of socialbots on twitter during the 1st 2016 us presidential debate. In *Proc. of the Int'l AAAI Conference on Web and Social Media (ICWSM)*, 2018.
- [80] Manuel Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 23–32, New York, NY, USA, 2013. ACM.
- [81] S. Romero. Dilma rousseff is ousted as brazil's president in impeachment vote. <https://www.nytimes.com/2016/09/01/world/americas/brazil-dilma-rousseff-impeached-removed-president.html>, 2016.
- [82] Avi Rosenfeld, Sigal Sina, David Sarne, Or Avidov, and Sarit Kraus. Whatsapp usage patterns and prediction models. *Demographic Research*, 39:647–670, 2018.
- [83] Amanda Rossi. Como o whatsapp mobilizou caminhoneiros, driblou governo e pode impactar eleições. <https://www.bbc.com/portuguese/brasil-44325458>, 2018.
- [84] Derek Ruths. The misinformation machine. *Science*, 363(6425):348–348, 2019.

- [85] Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 55–64. ACM, 2011.
- [86] Anika Schwind and Michael Seufert. Whatsanalyzer: a tool for collecting and analyzing whatsapp mobile messaging communication data. In *2018 30th International Teletraffic Congress (ITC 30)*, volume 1, pages 85–88. IEEE, 2018.
- [87] Michael Seufert, Tobias Hofffeld, Anika Schwind, Valentin Burger, and Phuoc Tran-Gia. Group-based communication in whatsapp. In *IFIP Networking Conference (IFIP Networking) and Workshops, 2016*, pages 536–541. IEEE, 2016.
- [88] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.
- [89] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *PloS one*, 13(4):e0196087, 2018.
- [90] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.
- [91] Sille Obelitz S oe. Algorithmic detection of misinformation and disinformation: Gricean perspectives. *Journal of Documentation*, 74(2):309–332, 2018.
- [92] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.
- [93] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. Ecosystem or echo-system? exploring content sharing across alternative media domains. In *Proc. of the Int’l AAAI Conference on Web and Social Media (ICWSM)*, 2018.
- [94] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.
- [95] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. In *Proc. of the Workshop on Data Science for Social Good (SoGood)*, 2017.

- [96] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.
- [97] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.
- [98] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.
- [99] Jeffrey E Thomas. Statements of fact, statements of opinion, and the first amendment. *Calif. L. Rev.*, 74:1001, 1986.
- [100] Amo Tong, Ding-Zhu Du, and Weili Wu. On misinformation containment in online social networks. In *Advances in Neural Information Processing Systems*, pages 339–349, 2018.
- [101] David Treece. Brazil’s ex-president lula imprisoned to keep him out of the election. <https://www.theguardian.com/world/2018/jun/08/brazils-ex-president-lula-imprisoned-to-keep-him-out-of-the-election-letters>, Jun 2018.
- [102] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):10, 2019.
- [103] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
- [104] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’18, pages 849–857, New York, NY, USA, 2018. ACM.
- [105] Shabeer Ahmad Wani, Sari M Rabah, Sara AlFadil, Nancy Dewanjee, and Yahya Najmi. Efficacy of communication amongst staff members at plastic and reconstructive surgery section using smartphone and mobile whatsapp. *Indian journal of plastic surgery: official publication of the Association of Plastic Surgeons of India*, 46(3):502, 2013.
- [106] Claire Wardle. The ‘trumpet of amplification’. <https://firstdraftnews.org/latest/5-lessons-for-reporting-in-an-age-of-disinformation/>, 2018.



- [107] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. The spread of medical fake news in social media—the pilot quantitative study. *Health Policy and Technology*, 7(2):115–118, 2018.
- [108] Martin Wattenberg and Fernanda B Viégas. The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics*, 14(6):1221–1228, 2008.
- [109] Duncan J Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 2004.
- [110] Yang Yang, Jie Tang, Cane Wing-ki Leung, Yizhou Sun, Qicong Chen, Juanzi Lit, and Qiang Yang. Rain: social role-aware information diffusion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 367–373. AAAI Press, 2015.
- [111] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, pages 405–417. ACM, 2017.
- [112] Li Zeng, Kate Starbird, and Emma S Spiro. Rumors at the speed of light? modeling the rate of rumor transmission during crisis. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1969–1978. IEEE, 2016.
- [113] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612. International World Wide Web Conferences Steering Committee, 2018.
- [114] H. Zhang, A. Kuhnle, J. D. Smith, and M. T. Thai. Fight under uncertainty: Restraining misinformation and pushing out the truth. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 266–273, Aug 2018.
- [115] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Emotional dynamics in the age of misinformation. *PloS one*, 10(9):e0138740, 2015.
- [116] Fabiana Zollo and Walter Quattrociocchi. Misinformation spreading on facebook. In *Complex Spreading Phenomena in Social Systems*, pages 177–196. Springer, 2018.

- 
- [117] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

# Appendix A

## Dictionary of words related to the 2018 Brazilian elections

In this Appendix, we present in Tables A.1 and A.2 and A.3 the list of words from a dictionary related to the 2018 Brazilian elections.

Presidents Candidates	Political parties	Governors Candidates	News
Lula	MDB	Carlos Gianazzi	Portal
Free Lula	PT	José Anibal	Chanel
Lula in Prison	PSDB	Celso Russomanno	News
Lula 2018	PP	Luiz Marinho	Blog
Lula President	PDT	João Doria	News
Bolsonaro 2018	PTB	Márcio França	Brazil News
Bolsonaro President	DEM	Anthony Garotinho	
Bolsonaro	PR	Índio da Costa	
Bolsomito	PSB	Celso Amorim	
Bolsominions	PPS	Leonardo Giordano	
Jair Bolosnaro	PSC	Eduardo Paes	
Geraldo Alckmin	PCdoB	Miro teixeira	
Alckmin President	PRB		
Alckmin 2018	PV		
Marina Silva	PSD		
Marina 2018	PRP		
Marina President	PSL		
Ciro Gomes	PHS		
Ciro 2018	PTC		
Ciro President	SD		
Aldo Rebelo	PSDC		
Aldo Rebelo president	AVANTE		
Aldo Rebelo 2018	PODE		
Manuela D'Avila	PSOL		
Manuela 2018	PRTB		
Manuela President	PROS		

Table A.1: Dictionary of words related to the 2018 Brazilian elections - Part 1

<b>Presidents Candidates</b>	<b>Political parties</b>
Álvaro Dias	PEN
Álvaro 2018	PPL
Álvaro President	PMB
Rodrigo Maia	PSTU
Rodrigo Presidente	PCB
Rodrigo 2018	Partido Novo
Michel Temer	PCO
Fora Temer	
Temer 2018	
João Amoedo Presidente	
Joao Amoedo 2018	
Guilherme Boulos	
Guilherme Boulos 2018	
Guilherme Boulos President	
Flavio Rocha	
Flavio Rocha President	
Flavio Rocha 2018	
Paulo Rabello	
Paulo Rabello 2018	
Paulo Rabello President	
Henrique Meirelles	
Henrique Meirelles President	
Henrique Meirelles 2018	
Fernando Collor	
Collor President	
Collor 2018	

Table A.2: Dictionary of words related to the 2018 Brazilian elections - Part 2

<b>Ideologies and Political Sides</b>	
Traditional Family	Feminism
Neoliberalism	Immigrant
Social Democrat	Military intervention
Communism	Jean Wyllys
Socialism	Kit Gay
Absolutism	Legalization of drugs
Landless Movement	Lei Rouanet
Marielle lives	Liberal
Moro	Freedom of expression
Corruption	Chauvinism
Lava Jato	Maria da penha
Senate	Maria do Rosário
Elections 2018	Marielle lives
Gay cure	Monarchy
It was not an accident	Moro
Carmen Lucia	Occupation
Petista	Patriot
Prison Aecio	Death penalty
We are many	Politics
Pretalhas	Reverse Racism
Coxinha	Skinhead
Activist	Conservatism
Abortion	Anarcho-capitalism
Anarchy	Ankara
Weapons	Tition
Family Government Program	Adultery
Capitalism	Feminazi (Feminist Girl)
Comunism	Feminists
Conservative	Sexism
Demilitarization of the police	Gays
Fascists	Fagots
Politically correct	Gender
Armament	Homophobia
Reduction of criminal age	Homosexual
Refugee	Lesbians
Socialist	LGBT
Go to Cuba	Pablllo Vittar
Venezuelan	Straight pride
Trans	Sexuality
Amazon Independence now	Trans
Independentism	Prison
Atheists	MST
Catholics	MTST
Believer	Nationalist
Creationism	UNE
God	Prison
Jewish	Nazism

Table A.3: Dictionary of words related to Ideologies and Political Sides