Analyzing Textual (Mis)Information Shared in WhatsApp Groups

Gustavo Resende
UFMG, Brazil
gustavo.jota@dcc.ufmg.br

Philipe Melo
UFMG, Brazil
philipe@dcc.ufmg.br

Julio C. S. Reis
UFMG, Brazil
julio.reis@dcc.ufmg.br

Marisa Vasconcelos
IBM Research
marisaav@br.ibm.com

Jussara M. Almeida
UFMG, Brazil
jussara@dcc.ufmg.br

Fabrício Benevenuto
UFMG, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

Whatsapp is a messenger app that is currently very popular around the world. With a user-friendly interface, it allows people to instantaneously exchange messages in a very intuitive and fluid way. The app also allows people to interact using group chats, sharing messages, videos, audios, and images. These groups can also be a fertile ground to spread rumors and misinformation. In this work, we analyzed the messages shared on a number of political-oriented WhatsApp groups, focusing on *textual content*, as it is the most shared media type. Our study relied on a dataset containing all textual messages shared in those groups during the 2018 Brazilian presidential campaign. 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. Our study aims at identifying characteristics that distinguish such messages from the other textual messages (with unchecked content). To that end, we analyzed various properties of the textual content (e.g., language usage, main topics and sentiment of message's content) and propagation dynamics of both sets of messages. Our analyses revealed that textual messages with misinformation tend to be concentrated on fewer topics, often carrying words related to the cognitive process of *insight*, which characterizes chain messages. We also found that their propagation process is much more viral with a distinct behavior: they tend to propagate faster within particular groups but take longer to cross group boundaries.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Applied computing** → *Sociology*.

KEYWORDS

misinformation, content dissemination, WhatsApp, textual information

1 INTRODUCTION

Whatsapp is a messenger app that 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. Moreover, the possibility of sending messages via the Internet, instead of using a text messaging service (e.g., SMS), is a much cheaper choice. One could argue that such features highly contributed to turning WhatsApp into the most popular messaging app in the world, with 1.5 billion users in 180 countries and 1 billion daily active users[1].

The conversation in groups allows users to chat and interact instantly with all of those who joined the group. The visibility of such conversation is restricted to members of the group, being thus controlled by the group manager who decides who can join the group. However, access to a group can be made effectively public when the manager shares the link to join it in websites or social networks like Facebook and Twitter. In such a case, anyone with access to the link can join the group, which can be considered, from a practical perspective, public.

WhatsApp groups facilitate the dissemination of different types of content including chain messages, news, memes and rumors, including the so-called fake news. We here are particularly interested in the spread of *misinformation*, which refers to reportedly false (or inaccurate) information which is often intended to deceive people. In countries like India and Brazil where the app reportedly already reached 200 and 120 million users respectively[2], the spread of misinformation in WhatsApp has had consequences for society such as lynching episodes in India[3] and fake news flooding during Brazilian presidential elections[4]. 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 in those two countries[5]. This is a first step to constrain the spread of fake news. Yet, given the great popularity of the application, its effectiveness is naturally limited. It is of utmost importance to identify characteristics of messages containing

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326029

[1] http://www.businessofapps.com/data/whatsapp-statistics/
[2] https://www.financialexpress.com/industry/technology/whatsapp-now-has-1-5-billion-monthly-active-users-200-million-users-in-india/1044468/
[3] https://www.livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html
[4] https://www.independent.co.uk/life-style/gadgets-and-tech/news/whatsapp-india-killings-latest-update-explained-app-fake-hoax-rumours-a8428746.html
[5] https://www.theguardian.com/technology/2019/jan/21/whatsapp-limits-message-forwarding-fight-fake-news

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 [5, 10] as well as its use on specific tasks (e.g., educational tasks, medical information exchange) [4, 32]. In a recent work [22], we have studied the dissemination of *images* in political public groups in WhatsApp, highlighting some differences in images containing previously identified misinformation from the rest. Yet, no prior study, not even ours, focused on exploring the presence of misinformation in *textual* messages, which is the most common type of content shared in the system, and whether there are particular features that distinguish them from the other textual messages.

Towards filling this gap, we here present an extensive investigation of the spread of textual messages within *WhatsApp public groups*. We focus on *political-oriented* public groups as we expect greater user engagement in topics of stronger social impact. We aim to compare the textual messages containing previously reported misinformation with other textual messages whose content was unchecked. We characterize these two sets of textual 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 tackle the following research questions.

RQ1: What are the differences in terms of textual features between messages containing misinformation and the rest?

RQ2: 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 textual messages?

Towards tackling these questions, we used the dataset collected in [22]. However, unlike in that work, which focused on image content, we here analyze the textual messages. The dataset was collected from 364 politically oriented groups. We joined those groups and gathered the content shared within them for the time period of the Brazilian first round of general elections campaign (August 16th to October 7th, 2018). We also 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 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). We also found that their propagation process is much more viral, reaching a larger number of users and groups, with a distinct behavior: they tend to propagate faster within particular groups but take longer to cross group boundaries, which results in such message lasting longer on the system.

The remainder of this paper is organized as follows. Next section discusses related work, while Section 3 describes the WhatsApp dataset used in our analysis. Sections 4 and 5 present our analyses of textual features and propagation dynamics, thus addressing RQ1 and RQ2, respectively. Finally, Section 6 concludes the paper and gives directions for future work.

2 RELATED WORK

A number of recent studies have investigated how online social networks may impact many global political scenarios, such as the White Helmets in the Syria[26] and the 2016 US presidential campaign. The latter attracted the focus of various studies, from the role of *bots* and political advocates on Twitter [23] to the influence of *fake news* on the results of the election [1, 9].

Some authors focused on the propagation of misinformation on the Web. For example, Castillo *et al.* [6] analyzed the credibility of information shared on Twitter, discovering that there are measurable differences in the way credible and not credible messages propagate, whereas Vosoughi *et al.* [30] showed that fake news tend to spread faster than the real news. Others have proposed learning methods to automatically detect fake messages ranging from lexical to deep learning approaches exploring linguistic and network features [25, 31]. On this classification task, textual features are great semantic resources used very often on many approaches using the language structure [6], sentiment and other psycho-linguistic cues [18, 30], topic models [11] and even political biases [2] of the messages. However, these prior efforts focused mostly on news articles and posts in online social networks such as Facebook [21], Twitter [6, 11] and Weibo [31]. None of them investigated the spread of misinformation on WhatsApp, which 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[6] where content dissemination is driven solely by the members' intentions, with no influence of any recommendation or news feed algorithm. Thus, information spread is such environment may convey particular properties worth studying.

Yet, there have been reports that WhatsApp is being massively used not only as an important tool for marketing[7], but also as a vehicle for spreading fake news. For example, Cunha *et al.* [7] pointed WhatsApp as one of the leading sources of misinformation spreading, showing how users are easily manipulated through the spread of misleading information.

A few recent studies have investigated how users behave as they share messages in WhatsApp, particularly within chat groups. Garimella *et al.* [10] proposed a generalizable data collection methodology for WhatsApp public groups, whereas Seufert *et al.* [24] investigated the emerging group-based communication paradigm on WhatsApp and its implications on mobile network traffic. Caetano *et al.* [5] in turn analyzed user behavior in public groups using a three-layer hierarchical approach (e.g., message, user and groups). Finally, some studies [4, 32] focused on understanding how users interact using Whatsapp for performing different tasks such as educational tasks, medical information exchange, etc. To our knowledge, the only prior work to tackle the dissemination of misinformation in WhatsApp is our own recent study [22]. Yet, our focus in that work was on the dissemination of image content in the system, analyzing images containing misinformation and proposing some general countermeasures. We have not analyzed textual messages shared in the groups, as we do here. Also, we

[6] The number of members in a group is limited to 256.
[7] http://nyti.ms/2L3AV3M

here do a more thorough comparison of the contents shared in the messages and their propagation dynamics, identifying features that clearly distinguish messages with misinformation from the rest.

In sum, while prior studies provide valuable knowledge about WhatsApp as an emerging social network and information dissemination vehicle, the analysis of misinformation in the system is still at a very early stage. This work greatly adds to the current literature by focusing on the textual properties and propagation dynamics of misinformation on WhatsApp.

3 WHATSAPP DATASET

As mentioned, this work relies on the same raw dataset of WhatsApp messages collected in our prior study [22]. Our present focus is on a large part of this dataset, covering the period of August 16th to October 7th, 2018 which precedes the first-round of the Brazilian general elections campaign when a new president, state governors, and House members were elected[8]. Moreover, unlike in [22], where our analyses were focused on image content, our present concern is on a different subset of the data, notably *textual* messages. In this section, we first present the filtered dataset used in this work (Section 3.1) and then describe our methodology to identify misinformation in the textual messages (Section 3.2).

As presented in [22], in order to gather a relevant WhatsApp dataset, we first identified a considerable number of public groups by searching for them using Google as well as Twitter and Facebook search engines. Specifically, we submitted the query "chat.whatsapp.com", a common piece of any URL containing an invite to join a WhatsApp group. 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[9]. This dictionary basically contains the name of all politicians, political parties, as well as words associated with political extremism. Finally, we performed a manual inspection of the collected group names to filter out those that are not related to politics. In total, we found 3,444 distinct links for public groups, out of which only 1,828 were valid. Due to memory restrictions, we were able to subscribe to 364 distinct groups, selected randomly out of the valid ones, using three cellphones we had available. After joining each group, the collection process was initiated.

WhatsApp uses end-to-end encryption, which makes access to the contents of the messages harder. Thus we used the *WebWhatsAppAPI*[10], which provides an interface in Python to send and receive messages by *WhatsApp Web* and uses *Sellenium* to automate the application through the browser. In total, we retrieved 591, 162 messages posted by 18, 725 unique users in the 364 groups. We emphasize that sensitive information such as user names and phone numbers were not stored in our dataset as we performed user anonymization before storing the data. Specifically, we mapped each telephone number to a unique user ID and we use these IDs to identify the source of each message[11]. For each message, we stored its content, the name of the group where the message was posted, group ID, user ID, date, and timestamp.

Table 1: Overview of dataset.

Period	08/16 – 10/7/2018
#Public Groups	364
#Total Users	18,725
#Textual Messages	591,162
#Images	110,954
#Videos	73,310
#Audios	14,488
Filtering: textual messages with > 180 characters	
#Textual Messages	59,979
#Distinct Textual Messages	37,674
# URLs	19,502

3.1 Our Filtered Dataset

Table 1 gives an overview of our collected dataset. Clearly, textual messages comprise the great majority on these groups, representing 75% of all messages shared. Since we aim at studying the presence of misinformation, we analyzed only messages with at least 180 characters 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.

We grouped similar content by computing the Jaccard similarity [8] 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 were grouped together and considered as (semi-)duplicates. The choice of threshold was made empirically. 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. In this process, a representative message for each content was selected, keeping information about the groups each content was sent to and their timestamps. In total, we identified 37, 674 distinct textual messages.

3.2 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 the messages in (filtered) dataset. Specifically, we crawled checked information (news or claims) from six popular Brazilian fact-checking sites: "Aos fatos", "Me engana que eu posto", "e-farsas", "é ou não é (G1)", "Lupa" and "Boatos.org"[12]. 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.

[8]The first round of elections occurred on October 7th. Most voted candidates who gathered more than 50% of the votes were automatically considered winners, which was the case of several candidates running for state governors. Otherwise, the two best-ranked candidates continued campaigning for a second round. This was the case of the presidential candidates.

[9]https://goo.gl/PdwAfV

[10]WhatsAppAPI available on http://github.com/mukulhase/WebWhatsapp-Wrapper

[11]Note that we are not able to identify individuals as the same person may have joined one or different groups using different telephone numbers.

[12]aosfatos.org, veja.abril.com.br/blog/me-engana-que-eu-posto/, www.e-farsas.com, g1.globo.com/e-ou-nao-e/, piaui.folha.uol.com.br/lupa/, and www.boatos.org

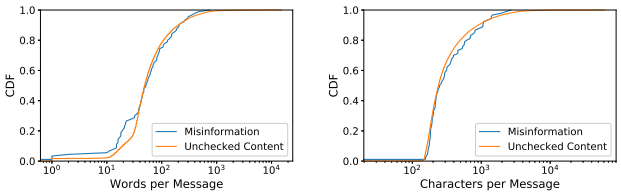
We then computed the text similarity between each textual message(filtered) in our WhatsApp dataset and each collected fact labeled as fake by at least one of the fact-checking websites. For the latter, we experimented with using only the contents of the summary field and using the description, which 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[13] to remove stop words and accents as well as to 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 [16]. 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 $cos(\boldsymbol{v}_m, \boldsymbol{v}_s) = \frac{\boldsymbol{v}_m \cdot \boldsymbol{v}_s}{||\boldsymbol{v}_m|| \cdot ||\boldsymbol{v}_s||}$. [14]

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 score, but containing retractions (e.g., links to fact-checker websites refuting the original content) were manually excluded from 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[15].

In the following sections, we compare properties of the textual content and propagation dynamics of the messages containing misinformation with the other textual messages in our dataset. We refer to the latter as *unchecked*. We do not claim that there is no misinformation in the unchecked messages, given that such an assertion is restricted to the availability of checked facts. Yet, we expect that we were able to catch most textual messages containing misinformation in our dataset, especially those with greater impact on users, as they most probably were reported by the fact checkers.

4 TEXTUAL PROPERTIES OF MESSAGE CONTENT

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 main topics and frequent words present in each type of message. Note that, compared to [22], all these analyses, which are focused on textual content, are novel.



(a) Number of words (b) Number of characters

Figure 1: Distributions of message sizes

4.1 Message Sizes

We start by looking at the sizes of the messages shared on WhatsApp. Figures 1(a) and 1(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. Recall that our filtered dataset contains only messages with more than 180 characters. According to Figure 1(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 have 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[16].

Figure 1(b) shows that both distributions of numbers of characters are very similar up to around 4,000 characters. 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, we do find some very long messages (up to 61,681 characters) among those with unchecked content.

4.2 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) [27]. LIWC is a psycholinguistic lexicon system that categorizes words into psychologically meaningful groups. We used the dictionary

[13] https://spacy.io/

[14] We did experiment with other similarity metrics, notably WMD (*Word Mover's Distance*)[13], which covers the semantics of sentences, and the results were similar, but with a higher cost of processing.

[15] Throughout this paper we use the term *sharing* 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.

[16] 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.

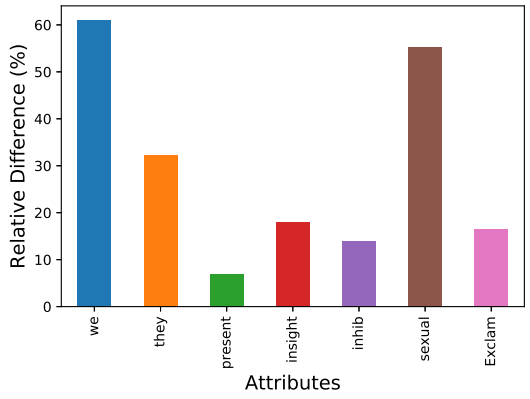


Figure 2: LIWC attributes that occur more frequently in messages with misinformation.

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 individual 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 [14], 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. We identified 7 (out of 92) attributes, for which the null hypothesis can be rejected with a confidence level of 0.95. We then computed the relative difference between the *average* values of each such attribute for messages with misinformation and messages with unchecked information.

Figure 2 shows the attributes with a greater presence in messages with misinformation. These are mainly subcategories of the linguistic attributes (*we, they, present, exclaim*) and psychological attributes (*insight, inhibition, sexual*).

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 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 tended to be 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.

4.3 Sentiment Analysis

Sentiment analysis has become an extremely popular tool to capture text polarity, especially on social media data [15]. In order to investigate the overall subjective cues of sentiment in the WhatsApp textual messages, we used a Portuguese version[17] of SentiStrength method [29] 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 [20]). We here employ it to investigate whether there are differences in the sentiment of messages carrying misinformation when compared to the rest.

Figure 3 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 [28]. However, the results in Figure 3 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.

[17] Available in: sentistrength.wlv.ac.uk.

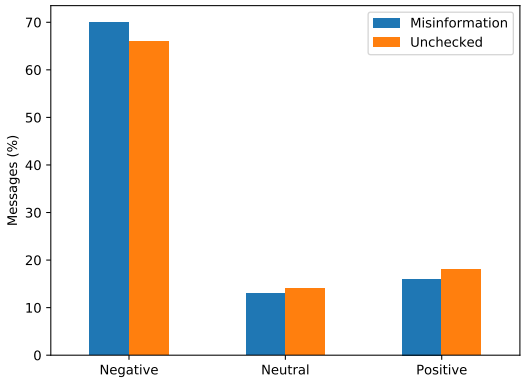


Figure 3: Sentiment polarity of messages.

In particular, messages with misinformation are slightly more negative. Such difference is indeed statistically significant by a Kruskal-Wallis H-test [12], with p-value < 0.005. This finding is in agreement with previous observations that misinformation content tends to be more negative [34], especially within polarized communities. Moreover, inspired by previous results on online news [20], 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.

4.4 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) [3], 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.

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* [19], 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 [17], 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 2, 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, characterize by words such as *God, life, money, millions* and *Facebook*.

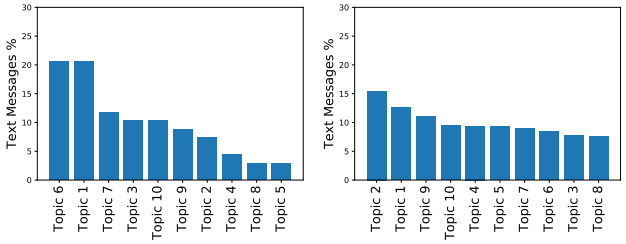
Table 2: Topics inferred by LDA algorithm.

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_whatsapp, 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(a) and 4(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 misnformation 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. The links present in the messages point to news about different candidates and polling surveys results. 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



(a) Messages with misinformation (b) Messages with unchecked content

Figure 4: Distributions of topics inferred by LDA.

candidates, influenced the 2018 presidential election campaign in Brazil[18]. 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.

4.5 Frequent Terms

To further support our analyses of the contents of the WhatsApp messages, Figure 5 shows the word clouds of the top 500 most frequent words (translated to English) for both sets of messages (with misinformation and unchecked content). These word clouds were produced using the Wordle tool[19]. 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) were 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 5(a) as input to the Word tree visualization tool [33], 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 6 shows one such word tree, rooted by the word *Please* (*Por favor*, in Portuguese). This was the root with the largest number of branches in the set of messages with misinformation. Indeed, as shown in Figure 6, 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 5(a), these phrases are related

(a) Messages with misinformation

(b) Messages with unchecked content

Figure 5: Word clouds of the top 500 words (translated to English).

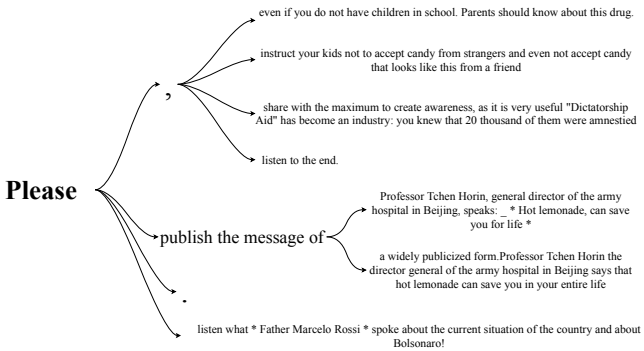


Figure 6: Word tree for the word root *Please*.

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 indicate 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 messages, being representative of the psychological process of *insight*. As found in Section 4.2, this psycholinguistic attribute does indeed occur more often in messages with misinformation.

[18] https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html?module=inline
[19] Available at: http://www.wordle.net/

5 PROPAGATION DYNAMICS

Besides characterizing the textual content of the messages, we also analyze their propagation dynamics within each group as well as across different groups. 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 in the system (Section 5.2). Compared to the results in [22], this provides a more detailed analysis of message propagation covering a larger number of metrics. Nevertheless, in the following, we explicitly compare our present findings with those in our prior work (despite the focus on different media types) for metrics analyzed in both studies.

5.1 Message Reach

Recall that, as presented in Section 3.1, we do group 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 7(a), 7(b) and 7(c) show the cumulative distributions of those measures for messages containing misinformation as well as messages with unchecked content.

As shown in Figure 7(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). Similarly, the messages with misinformation tend to reach a much larger number of distinct groups. Figure 7(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 7(c), the distinction between messages with misinformation and with unchecked content is similarly very drastic when it comes to the total number of shares. 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, textual messages with misinformation have a much greater reach in WhatsApp, suggesting a viral behavior within and across the WhatsApp groups.

5.2 Temporal Properties of Message Spread

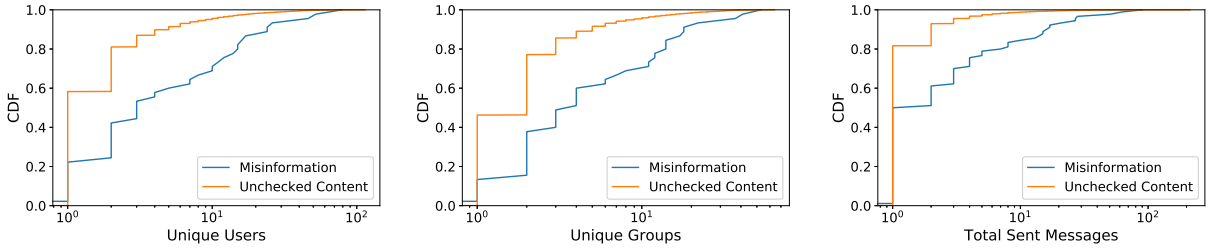
In this section, we investigate the spread of a message over time, focusing on messages that were shared at least twice in our dataset 50% and 20% of the messages with misinformation and with unchecked content, respectively). We analyze temporal properties of this spread including message *lifetime* and the time between consecutive shares of the same message, here referred to as *burst time*. These two metrics were also analyzed in [22] for images, and we compare our results with those prior findings below. 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*).

5.2.1 Message 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 8(a) shows the cumulative distributions of the lifetimes (in terms of days) for each set of messages. Clearly, the messages with misinformation tend to remain in the system for much longer: roughly half of the messages with misinformation in our dataset had a lifetime of at least 10 days. In contrast, most 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. This result is in sharp contrast with our prior findings for images shared in WhatsApp groups[22], where we did *not* observe clear differences between the lifetimes of images with misinformation and 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. Figure 8(b) shows the cumulative distributions of burst time for messages with misinformation and unchecked content. We note that the two distributions 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 messages with misinformation. That is, unchecked content is 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. 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.

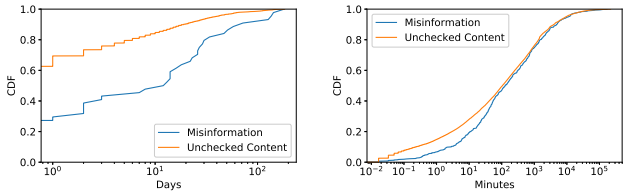
Once again, we observe differences with our previous findings for image content. In [22], we reported the opposite pattern, with shorter burst times for messages with misinformation. Recall that the two studies rely on data collected from the same WhatsApp groups during the same period, but focus on different types of messages. Thus, the misinformation propagation patterns does 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. As our final analysis, we 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



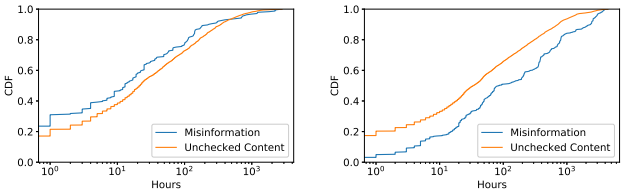
(a) Number of unique users per message (b) Number of unique groups per message (c) Total number of shares

Figure 7: Cumulative distributions of the reach of each message in terms of distinct users, distinct groups and total shares.



(a) Message Lifetimes (b) Message Burst Times

Figure 8: Cumulative distributions of life and burst times.



(a) Intra-group times (b) Inter-group times

Figure 9: Cumulative distributions of inter and intra-group times.

and inter-group 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. Note that the same message may be shared multiple times in the same group by one or more users. This metric is computed for messages within each group separately, and is restricted to messages shared at least twice in the group 29% and 9% of messages with misinformation and unchecked content, respectively. 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 83% and 55% of messages with misinformation and unchecked content, respectively.

We note that the analysis of intra and inter times may help understanding 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 are shown in Figures 9(a) and 9(b), respectively. The distributions for messages with misinformation and with unchecked content are different. Within each group, the shares of messages with misinformation tend to be somewhat more concentrated in time, happening faster. As Figure 9(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%. In contrast, Figure 9(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. In contrast, for messages with unchecked content, 30% of the inter-group times are within the same limit.

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

6 CONCLUDING REMARKS

We have analyzed textual properties and propagation dynamics of messages disseminated in a number of political-oriented WhatsApp groups. Our study was driven by the goal of identifying properties that distinguishing textual messages containing misinformation from other textual messages whose content is unchecked. 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 by means of textual similarity analysis.

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 is 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. 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 [34].

Our analyses of 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. Moreover, 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. Interestingly, these results are in contrast with our prior study of misinformation in *images* shared in WhatsApp [22], suggesting that the propagation dynamics of misinformation may indeed depend on the type of media used to convey the information.

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, monitoring period), they might generalize, to some extent, to other WhatsApp groups and time period. 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 on WhatsApp. We hope it motivates follow-up efforts covering other datasets, time periods, WhatsApp groups and media types. Exploring the analyzed features in the design of automatic mechanisms for detecting misinformation on WhatsApp is also a promising future work.

ACKNOWLEDGMENTS

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-1, as well as grants from Google, CNPq, CAPES, and Fapemig.

REFERENCES

[1] H. Allcott and M. Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.
[2] Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Fabricio Benevenuto, Krishna P. Gummadi, and Adrian Weller. 2018. Purple Feed: Identifying High Consensus News Posts on Social Media. In *Proc. of the Conf. on Artifical Intelligence, Ethics Society.*
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] Dan Bouhnik and Mor Deshen. 2014. WhatsApp goes to school: Mobile instant messaging between teachers and students. *Journal of Information Technology Education: Research* 13, 1 (2014), 217–231.
[5] Josemar Alves Caetano, Jaqueline Faria de Oliveira, Hélder Seixas Lima, Humberto T Marques-Neto, Gabriel Magno, Wagner Meira Jr, and Virgílio AF Almeida. 2018. Analyzing and characterizing political discussions in WhatsApp public groups. *arXiv preprint arXiv:1804.00397* (2018).
[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proc. of the Int'l Conference on World Wide Web.*
[7] Evandro Cunha, Gabriel Magno, Josemar Caetano, Douglas Teixeira, and Virgilio Almeida. 2018. Fake news as we feel it: perception and conceptualization of the term "fake news" in the media. In *Proc. of the Int'l Conference on Social Informatics.*
[8] Société Vaudoise des Sciences Naturelles. 1864. *Bulletin de la Société vaudoise des sciences naturelles.* Vol. 7. F. Rouge.
[9] A. Fourney, M. Racz, G. Ranade, M. Mobius, and E. Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proc. of the Int'l Conference on Information and Knowledge Management.*
[10] K. Garimella and G. Tyson. 2018. WhatsApp, Doc? A First Look at WhatsApp Public Group Data. In *Proc. of the Int'l Conference on Web and Social Media.*
[11] Jun Ito, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama. 2015. Assessment of tweet credibility with LDA features. In *Proc. of the Int'l Conference on World Wide Web.*
[12] H Kruskal and W Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.
[13] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning.* 957–966.
[14] F. Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.
[15] Philipe F Melo, Daniel H Dalip, Manoel M Junior, Marcos A Gonçalves, and Fabrício Benevenuto. 2019. 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).
[16] Ashish Moon and T Raju. 2013. A survey on document clustering with similarity measures. *International Journal of Advanced Research in Computer Science and Software Engineering* 3, 11 (2013), 599–601.
[17] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.* ACL.
[18] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. 2011. Detecting and tracking political abuse in social media. In *Proc. of the Int'l Conference on Web and Social Media.*
[19] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proc. of the Workshop on New Challenges for NLP Frameworks.*
[20] Julio Reis, Fabricio Benevenuto, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the News: First Impressions Matter on Online News. In *Proc. of the Int'l Conference on Web and Social Media.*
[21] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019).
[22] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proc. of the Web Conference.*
[23] Marian-Andrei Rizoiu, Timothy Graham, Rui Shang, Yifei Zhang, Robert Ackland, Lexing Xie, et al. 2018. # DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 US Presidential Debate. In *Proc. of the Int'l Conference on Web and Social Media.*
[24] Michael Seufert, Tobias Hoßfeld, Anika Schwind, Valentin Burger, and Phuoc Tran-Gia. 2016. Group-based communication in WhatsApp. In *Proc. of the IFIP Networking Conference and Workshops.* IEEE.
[25] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[26] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains. In *Proc. of the Int'l Conference on Web and Social Media.*
[27] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[28] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* 62, 2 (2011), 406–418.
[29] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.
[30] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[31] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proc. of the Int'l Conference on Knowledge Discovery & Data Mining.*
[32] Shabeer Ahmad Wani, Sari M Rabah, Sara AlFadil, Nancy Dewanjee, and Yahya Najmi. 2013. 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 (2013), 502.
[33] Martin Wattenberg and Fernanda B Viégas. 2008. The word tree, an interactive visual concordance. *Transactions on visualization and computer graphics* 14, 6 (2008), 1221–1228.
[34] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Emotional dynamics in the age of misinformation. *PloS one* 10, 9 (2015), e0138740.