COVID19.BR: A Dataset of Misinformation about COVID-19 in Brazilian Portuguese WhatsApp Messages

Antônio Diogo Forte Martins[1], Lucas Cabral[1], Pedro Jorge Chaves Mourão[2], Ivandro Claudino de Sá[1], José Maria Monteiro[1], Javam Machado[1]

[1]Department of Computing, Federal University of Ceará, Fortaleza-Ceará, Brazil

[2]Universidade Estadual do Ceará, Fortaleza-Ceará, Brazil

{diogo.martins, jose.monteiro, javam.machado}@lsbd.ufc.br

{lucascabral, ivandroclaudino}@aridalab.dc.ufc.br

pedro.mourao@aluno.uece.br

Abstract. *Nowadays, our society suffers with a major issue that unfortunately is becoming more and more problematic, once again through social networks, that is the misinformation. The primary source of misinformation in Brazil is the messaging application WhatsApp. However, due to WhatsApp's private messaging nature, there still few misinformation data sets built specifically from this platform. In this context, building a data set of WhatsApp messages about COVID-19 in Brazilian Portuguese and label misinformation messages within it becomes a crucial challenge. In this work, we present the COVID-19.BR, a data set of WhatsApp messages about coronavirus in Brazilian Portuguese, collected from Brazilian public groups and manually labeled.*

1. Introduction

Nowadays, our society suffers with a major issue that unfortunately is becoming more and more problematic, once again through social networks, that is the misinformation. In April 2020, the United Nation (UN) reported that we are living in a 'dangerous misinformation epidemic" responsible for the spread of misleading solutions about the coronavirus [1]. In February 2020, the Brazilian Health Ministry reported that among 6,500 messages received and analyzed by it, between January 22 and February 27, 90% were related to the new virus. From the messages about coronavirus, 85% were false[2].

The misinformation concept can be defined as a process of intentional production of a communicational environment based on false, misleading, or decontextualized information to cause a communicational disorder [Su et al. 2020]. Nevertheless, the term fake news, despite specifically describe intentionally misleading information written as journalistic news, has become very present in popular culture and is sometimes used as a misinformation synonym [Guo et al. 2019].

Currently, misinformation spreaders preferred distribution channel is WhatsApp instant messaging application. WhatsApp is very popular in Brazil, it has more than 120

[1]UN. "Hatred going viral in 'dangerous epidemic of misinformation' during COVID-19 pandemic". 14 April, 2020. Available in: https://news.un.org/en/story/2020/04/1061682. Accessed on: April 25, 2020.

[2]Available in: https://www.saude.gov.br/fakenews. Accessed in: April 25, 2020

million users in Brazil [Resende et al. 2019]. The Panorama Mobile Time/Opinion Box survey from February 2020 on Brazilian mobile messaging revealed that WhatsApp is installed on 99% of Brazilian smartphones. Among application users, 98% said they access it every day or almost every day [3]. A survey by the Oswaldo Cruz Foundation (Fiocruz) showed that 73.7% of the false news about the new coronavirus circulated through WhatsApp. Another 10.5% were published on Instagram and 15.8% on Facebook[4].

WhatsApp have a very relevant and interesting feature which is the possibility to join public groups. They are accessible through invitation links available on popular websites and social networks. Typically, they have specific discussion topics such as health, politics, and sports. Each group can have up to 256 members. So, WhatsApp public groups are very comparable to social networks. Consequently, they have been used to disseminate misinformation. Due to the volume of information we are exposed, we can not quickly distinguish what is or what is not misinformation [Vosoughi et al. 2018, Qiu et al. 2017].

Despite the scientific community's efforts, there is still a need for a large-scale corpus containing WhatsApp messages in Portuguese about COVID-19. In this sense, we provide COVID19.BR, a large-scale, labeled, anonymized, and public data set formed by WhatsApp messages in Brazilian Portuguese (PT-BR) about coronavirus pandemic, collected from public WhatsApp groups using the platform proposed in [de Sá et al. 2021]. With this data set, researchers will be able to build models to perform automatic misinformation detection, understand which parts of a message are important to detect misinformation, and understand how these misinformation messages spread through public groups.

2. Related Work

In [Gaglani et al. 2020], the authors contextualize the problem of spreading fake news on WhatsApp, especially in India and Brazil, and proposes a strategy for the automatic detection of fake news. A total of 10 public groups were scrapped for one week to get 1000 multilingual messages. After cleaning the data, the multilingual data was translated into English by employing the google translate API.

In [Resende et al. 2018], the authors presented a system for gathering, analyzing, and visualize public groups in WhatsApp. Besides describing their methodology, the authors also provide a brief characterization of the 169.154 messages shared by 6,314 users in 127 public groups to help journalists and researchers understand the repercussion of events related to the 2018 Brazilian elections. In [de Sá et al. 2021], the authors presented a platform, called Digital Lighthouse, for finding, gathering, analyzing, and visualize public groups in WhatsApp.

In the study presented in [Machado et al. 2019], the authors collected and analyzed 298,892 WhatsApp' messages, from 130 public groups, in the period leading up to the two rounds of the 2018 Brazilian presidential elections. Further, they examined a sam-

[3]SCHERMANN, Daniela. Panorama Mobile Time/Opinion Box: Mensageria no Brasil. Opinion Box, 2 mar. 2018. Available in https://blog.opinionbox.com/mensageria-no-brasil-sexta-edicao/. Accessed in: 11 mar. 2020.

[4]Available in: https://portal.fiocruz.br/noticia/pesquisa-revela-dados-sobre-fake-news-relacionadas-covid-19. Accessed in: 27 April, 2020.

ple of 200 videos and images extracted from these WhatsApp messages and developed a new typology to classify this media content.

In [Resende et al. 2019], the authors analyzed different aspects of WhatsApp messages from public political-oriented groups. The messages were collected during major social events in Brazil: a national truck drivers' strike and the Brazilian presidential campaign. The authors analyzed the types of content shared within such groups and the network structures that emerge from user interactions. Besides, they identified misinformation among the shared images using labels provided by journalists and by an automatic procedure based on Google searches. However, none of these works provides an entire public platform for finding, gathering, analyzing, and visualizing public groups in WhatsApp.

The FakeWhatsApp.BR, a dataset of WhatsApp messages in Brazilian Portuguese about the 2018 Brazilian elections, collected from public groups and manually labeled was presented in [Cabral et al. 2021]. Besides, the authors evaluated a series of misinformation classifiers combining Natural Language Processing-based techniques of feature extraction and a set of well-know machine learning algorithms, totaling 108 different scenarios. Their best result achieved a F1 score of 0.73.

3. Application

In [Martins et al. 2021], the authors used COVID19.BR to investigate different machine learning methods in order to build an efficient automated misinformation detection (MID) model for WhatsApp messages. They achieved an F1 score of 0.774 due to the predominance of short texts. However, when texts with less than 50 words are filtered, the F1 score rises to 0.85.

However, there are several applications for this data set. From the misinformation classification perspective, explore deep learning architectures in order to improve MID performance. From a qualitative point of view, characterize the messages by analyzing its content, explain how the messages spread through the groups, and use eXplainable Artificial Intelligence (XAI) to understand how the MID models are working.

4. Data Set Design

Despite the scientific community's efforts, there is still a need for a large-scale corpus containing WhatsApp messages in Portuguese about COVID-19. In this sense, we designed a large-scale labeled corpus of WhatsApp messages in Brazilian Portuguese inspired by [Silva et al. 2020].

In [Rubin et al. 2015], the authors suggest a methodological guideline for building corpora of deceptive content, and this guideline states that: the corpus must have truthful and their corresponding untruthful text versions, to allow finding patterns in instances of both labels; the texts should be in plain text format; the texts should have similar sizes to avoid bias in learning; the texts should belong to a specific time interval, because writing styles may change over time; and the corpus should keep the related metadata, such as the news URL and the authors because it can be useful for the fact-checking algorithms.

4.1. Collecting Data

Due to its private chat purpose, WhatsApp does not provide a public API to automatically collect data, different from other social media that the data is public. For this reason, the design process of a data set of WhatsApp messages is a technical, also ethical, challenge. We took a methodology similar to [Garimella and Tyson 2018, Resende et al. 2018].

We collected 228061 messages from open WhatsApp groups. These groups were found by searching for "`chat.whatsapp.com/`" on the Web and manually analyzing its theme and purpose. We only joined groups with at least 100 members to explore only relevant and active content groups. After analyzing and select, we were able to join 236 public groups. Then, we had to create a WhatsApp account to join the selected groups and collect the messages and their metadata. The users did not know we were collecting the messages. We collected messages between April and June 2020. After this period, we built a data matrix where each row corresponds to a group's message. The matrix columns are date, hour, phone number, international phone code, if the user is Brazilian its state, the text content of the message, word count, character count, and if the message contained media (audio, image, or video). Since our goal is to identify misinformation in the text of a WhatsApp message, our data set does not contain media files. Furthermore, we also counted the number of appearances of the same message in the data set. For this task, we only consider identical textual content from messages with more than five words. With this strategy, we could filter common messages like greetings. These messages that appear more than once in our data set, we call them "viral messages".

4.2. Data Anonymization

We tackle users' privacy issues by anonymizing their names and call phone numbers. Using a hash function, we created a unique and anonymous identifier for each user using the cell phone number as input. We set an alias for each group to achieve their anonymization. Since these groups are publicly available, our approach does not violate WhatsApp's privacy policy[5].

4.3. Labeling Process

Data labeling is another hard challenge since we have to specify if the text is true or false based on trusted sources, such as specialized journalists or fact-checking sites.

Hereafter, we describe our WhatsApp messages' textual content labeling process. In order to build a high-quality corpus, we conducted the data labeling process entirely manually. A human specialist checked each message's content and determined if it contains or not misinformation. Since this process is time-consuming, we chose to label only unique messages containing the following keywords: *"covid", "coron", "virus", "china", "chines", "cloroquin", "vacina"*. The resulting data set now has 2899 unique messages. This subset contains various types of messages such as fake news, rumors, true news, opinions, jokes, and hate speech. We labeled all these messages with the general misinformation definition adopted in [Su et al. 2020] labeling them as 0 if the message does not contain misinformation and 1 if it contains misinformation. Three annotators, two computer science masters students and one sociologist, conducted the labeling process. We solved labeling disagreements executing a collective review round.

[5]https://www.whatsapp.com/legal/privacy-policy

Our labeling process guideline is based on the following items:

1. If the message text content contains verifiable untrue claims, we annotate it as misinformation. We made use of trustful Brazilian fact-checking platforms such as *Agência Lupa*[6] and *Boatos.org*[7].
2. If the message text content contains imprecise, biased, alarmist, or harmful claims that cannot be proven, we annotate it as misinformation.
3. If the message text content is short and accompanied by media content (image, video, or audio), we search on the web for the media content and, if we find the corresponding media, we decide the label based on the previous criteria. If the original media cannot be found, we use the Item 2 criterion to label it.
4. If none of the previous criteria is found in the message text content, we label it as not containing misinformation.

After the labeling process, we removed the messages that have less than five words, messages not related to the coronavirus pandemics, messages containing only daily news summaries, and messages with only *url* as text content. The resulting corpus contains 865 unique messages labeled as misinformation (label 1) and 1178 unique messages labeled as non-misinformation (label 0).

Table 1 presents basic statistics about the corpus, including some traditional NLP features based on the number of tokens, types, characters, as well as the average number of shares, i.e., the frequency of the message in the original data set. In this new data set, we have a class ratio of 1.36 pointing to a slight class imbalance. Misinformation messages in average have more words and their length tend to vary more than regular messages which means that misinformation is spread in different writing styles. Both classes have similar number of shares.

Table 1. Data set basic statistics.

Statistics	Non-misinformation	Misinformation
Count of unique messages	1178	865
Mean and std. dev. of number of tokens in messages	82.80 ± 57.59	169.72 ± 243.76
Minimum number of tokens	5	5
Median number of tokens	22	52
Maximum number of tokens	2210	1666
Mean and std. dev. of number of types in messages	57.59 ± 96.59	109.03 ± 133.15
Average size of words (in characters)	5.82	5.12
Type-token ratio	0.696	0.642
Mean and std. dev. of shares	2.02 ± 4.17	1.89 ± 2.76

5. Data Set Description

COVID19.BR is provided in a *csv* file where the columns are date, hour, phone number, international phone code, if the user is Brazilian its state, the text content of the message,

[6] http://piaui.folha.uol.com.br/lupa/
[7] http://www.boatos.org/

	id	date	hour	ddi	country	country_iso3	ddd	state	group	midia	url	characters	words	viral	sharings	text
0	146759200457638065	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	9	1	0	1	Morreram?
1	146759200457638065	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	24	4	0	1	Olá novato, se apresente
2	5788106393468158140	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	9	2	0	1	há tempos
3	146759200457638065	07/04/20	04:07	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	13	2	0	1	Legião Urbana
4	5788106393468158140	07/04/20	04:13	55	BRASIL	BRA	21	Rio de Janeiro	2020_1	0	0	6	1	0	1	Indios

Figure 1. Extract from the collected data before the labeling process.

word count, character count, and if the message contained media (audio, image, or video). Each row represents a WhatsApp message. Figure 1 shows an extract from our data set after anonymization and before data labeling. We considered the complete data set to build the following visualizations.

In Figure 2, we have the proportion between the number of messages containing only text and messages containing media such as photos, videos or audios. The majority of the messages, 67.1%, in the data set contains only textual content.

In Figure 3, we have the proportion between the number of messages containing *url* or not. We can observe that 90.67% of the messages do not contain *url* in its content. Messages containing *urls* are important in the context of misinformation because these links can link the user to deceptive content in other platforms.

In Figure 4, we have the proportion between viral and non-viral messages, accordingly to our definition of a viral message. Only 6.38% of the messages are viral under our definition.

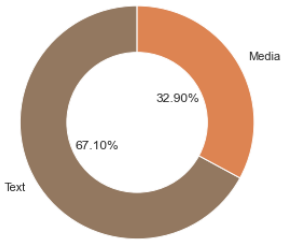

Figure 2. Messages with only text X containing media.

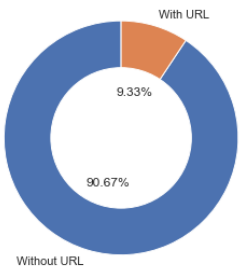

Figure 3. Messages containing or not urls.

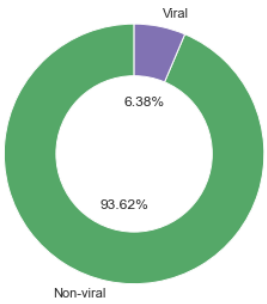

Figure 4. Non-viral X viral messages.

In Figure 5, we have the number of messages per Brazilian state. São Paulo leads the ranking of messages sent and captured in our data set. Minas Gerais, Rio de Janeiro, Bahia and Ceará also figure into the top 5 states with most messages sent. Amapá is the state with less messages in our data set.

In Figure 5, we have the number of users per Brazilian state. São Paulo leads the ranking again explaining with it is the state with most messages. Minas Gerais, Rio de Janeiro, Bahia and Ceará are again into the top 5 states. Amapá is the state with less user in our data set as well. Number of messages is a reflex of the number of users.

In Figure 7, we have the ratio between number of users and number of messages per state. Although Mato Grosso do Sul is the number 17 on the rank of number of messages and users, their users are the most activity among the other states. Only Minas Gerais is still figuring in the top 5 most active states.

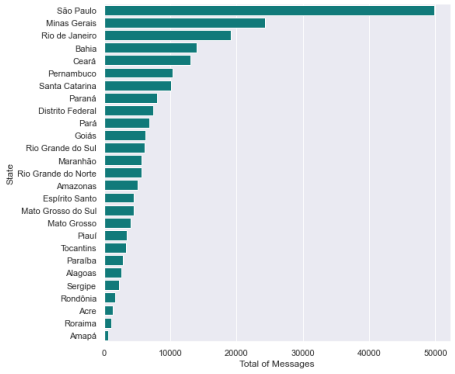

Figure 5. Number of messages per state.

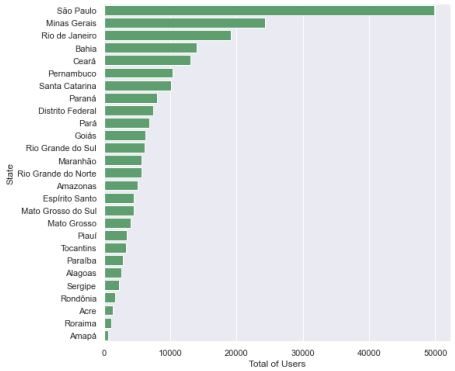

Figure 6. Number of users per state.

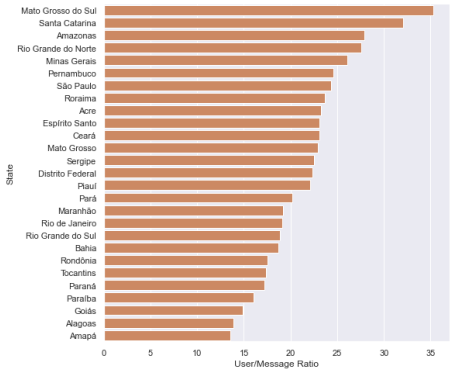

Figure 7. User/message ratio per state.

In Figure 8, we have the proportion between the number of Brazilian and not-Brazilian phone numbers. Only 1.77% of the users phone numbers are not from Brazil. So far we have researched, these non-Brazilian users are probably bots, but of course there are Brazilians living all over the world and keeping contact with Brazil through WhatsApp groups.

In Figure 9, we have the number of messages per country, excluding Brazil. Portugal is the country with the most number of messages followed by the United States.

In Figure 10, we have the distribution of message shares. We can observe that the most of the messages are shared 2 times with another peak at 21 shares.

In Figure 11, we have a wordcloud visualization. We can observe the words that appear the most in the messages. The words *covid*, *Brasil*, *coronavirus*, *saúde*, and *pessoa* are highlighted in this visualization, with this we can infer that the pandemics is being highly discussed in the groups.

In Figure 12, we have a wordnet visualization of the most frequent bigrams in the

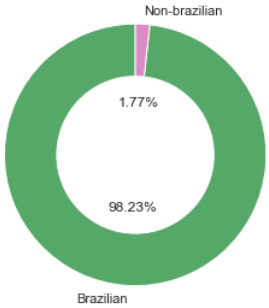

Figure 8. Proportion of messages from brazilian and not-brazilian numbers.

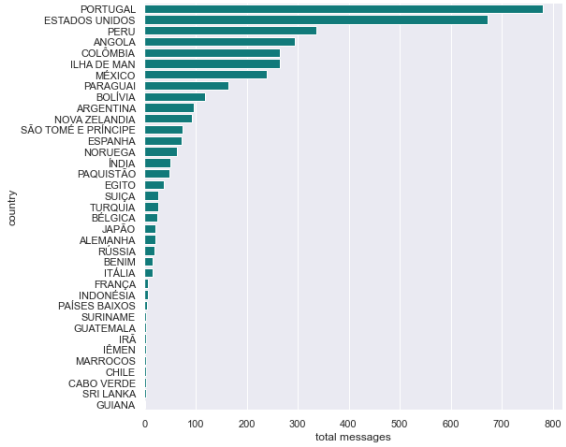

Figure 9. Number of messages per country.

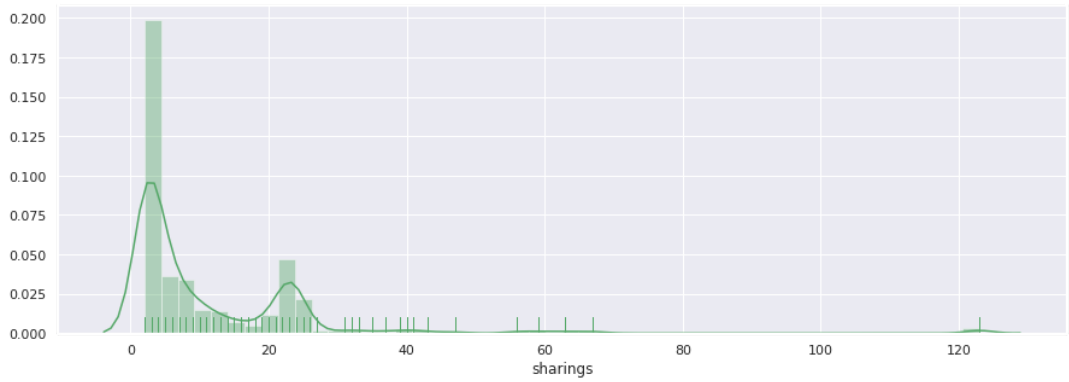

Figure 10. Distribution of message shares.

data set. We can observe that there are three main clusters of words. One is about the deaths caused by the new coronavirus pandemics. The other is the users talking about the various aids Brazilian parliamentarians receive. And the last and bigger one is actually a misinformation highly shared in the groups.

6. Conclusion

Actually, the large-scale dissemination of misinformation through social media has become a critical issue, harming social stability, democracy, and public health. In Brazil,

Figure 11. Corpus wordcloud.

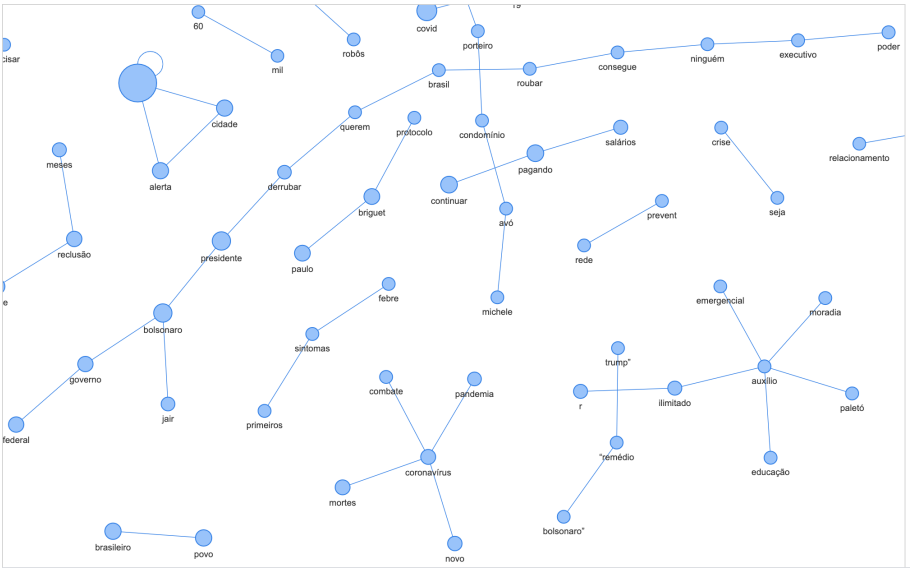

Figure 12. Corpus wordnet.

48% of the population use WhatsApp to get news. So, many groups have been used this instant messaging application in order to spread misinformation, especially as part of articulated political or ideological campaigns. In this work, we presented a large-scale, labeled, and public data set of WhatsApp messages in Brazilian Portuguese about coronavirus pandemic. We hope that COVID19.BR data set, available at `https://bityli.com/12KBp` (DOI: 10.5281/zenodo.5193932), can help journalists and researchers to understand the misinformation propagation in Brazil.

References

Cabral, L., Monteiro, J. M., da Silva, J. W. F., Mattos, C. L. C., and Mourão, P. J. C. (2021). Fakewhastapp.br: NLP and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*, pages 63–74. SCITEPRESS.

de Sá, I. C., Monteiro, J. M., da Silva, J. W. F., Medeiros, L. M., Mourão, P. J. C., and da Cunha, L. C. C. (2021). Digital lighthouse: A platform for monitoring public groups in whatsapp. In *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*, pages 297–304. SCITEPRESS.

Gaglani, J., Gandhi, Y., Gogate, S., and Halbe, A. (2020). Unsupervised whatsapp fake news detection using semantic search. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 285–289. IEEE.

Garimella, K. and Tyson, G. (2018). Whatsapp, doc? a first look at whatsapp public group data. *arXiv preprint arXiv:1804.01473*.

Guo, B., Ding, Y., Yao, L., Liang, Y., and Yu, Z. (2019). The future of misinformation detection: New perspectives and trends.

Machado, C., Kira, B., Narayanan, V., Kollanyi, B., and Howard, P. (2019). A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. WWW '19, page 1013–1019, New York, NY, USA. Association for Computing Machinery.

Martins, A. D. F., Cabral, L., Chaves Mourão, P. J., Monteiro, J. M., and Machado, J. (2021). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *Natural Language Processing and Information Systems*, pages 199–206, Cham. Springer International Publishing.

Qiu, X., Oliveira, D. F., Shirazi, A. S., Flammini, A., and Menczer, F. (2017). Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1(7):0132.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures.

Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., and Benevenuto, F. (2018). A system for monitoring public political groups in whatsapp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, page 387–390, New York, NY, USA. Association for Computing Machinery.

Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Su, Q., Wan, M., Liu, X., and Huang, C.-R. (2020). Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research*, 1:1–13.

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