Digital Lighthouse Platform: Understanding the Misinformation Phenomenon on WhatsApp

José Maria da Silva Monteiro Filho[1]
Ivandro Claudino de Sá[1]
Lucas Cabral Carneiro da Cunha[1]
Helena Martins do Rego Barreto[2]
Pedro Jorge Chaves Mourão[3]

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brazil
http://ufc.br

monteiro@dc.ufc.br,ivandro.claudino@alu.ufc.br,lucascabral@aridalab.dc.ufc.br

[2]Instituto de Cultura e Arte – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brazil
http://ufc.br

helena.martins@ufc.br

[3]Pós-Graduação em Sociologia – Universidade Estadual do Ceará (UECE)
Fortaleza – CE – Brazil
http://uece.br

pjmourao_cs@hotmail.com

Abstract. *In the past few years, the large-scale dissemination of misinformation through social media has become a critical issue, harming the trustworthiness of legit information, social stability, democracy and public health. In many developing countries such as Brazil, India, and Mexico, one of the primary sources of misinformation is the messaging application WhatsApp. In February 2020, the Panorama Mobile Time/Opinion Box survey on mobile messaging in Brazil revealed that WhatsApp was installed on 99% of Brazilian smartphones. Among users of the application, 98% said they access it every day or almost every day. In this context, WhatsApp provides an important feature: the public groups. Many of these groups have been used to spread misinformation, especially as part of articulated political or ideological campaigns. Despite this scenario, due to WhatsApp's private messaging nature, few methods were explicitly developed to investigate the misinformation phenomenon on this platform. This tutorial provides an overview of recent developments in monitoring misinformation spreading, automatic misinformation detection, and identifying misinformation spreaders. In addition, we provide an overview of the leading open problems associated with the misinformation phenomenon and briefly examine some of the existing solutions. We hope that our tutorial can help researchers better understand Brazil's misinformation propagation and use data science methods to face this critical phenomenon.*

1. Introduction

In the last years, the popularity of instant messaging applications has contributed to the spread of misinformation. Such applications allow content to be spread without editorial judgment. Through these systems, misinformation can deceive thousands of people in a short time (due to their appealing nature) and cause significant harm to individuals or society. Misinformation spreads faster, deeper, and broader in social media than legit information. Further, due to the high volume of information that we are exposed to when using social media, humans have a limited ability to distinguish true information from misinformation [Vosoughi et al. 2018]. In this context, misinformation has been used with malicious intentions to manipulate public opinion, change political scenarios, spread diseases, harm the democracy, individuals, organizations, or social groups, and obtain economic or political gains [Vosoughi et al. 2018, Guo et al. 2019, Su et al. 2020].

It is important to highlight that misinformation is a wide concept that can be defined in a general way as misrepresented information, including fabricated, misleading, false, fake, deceptive, or distorted information [Su et al. 2020]. This broad definition covers a variety of concepts such as fake news, rumor, deception and hoaxes. However, despite describing intentionally misleading information written as journalistic news, the term fake news has become very present in popular culture. It sometimes is used as a misinformation synonym [Guo et al. 2019].

The WhatsApp instant messaging application is very popular in Brazil, with more than 120 million users in about 210 million people. WhatsApp makes it possible to instantly share different media types, such as images, audio, and videos. In December 2019, a survey carried out by the Brazilian Chamber of Deputies and the Senate with 2,400 people concluded that 79% of people use the WhastApp as the main source of information[1]. In February 2020, the Panorama Mobile Time/Opinion Box survey on mobile messaging in Brazil revealed that WhatsApp is installed on 99% of Brazilian smartphones. Among users of the application, 98% said they access it every day or almost every day [2].

WhatsApp provides an essential feature: the public groups. These public groups are accessible through invitation links published on popular websites and various social networks, such as Facebook and Twitter. Usually, they have specific topics for discussion, such as politics, soccer, and education. WhatsApp allows each public group to have a maximum of 256 members. In this way, WhatsApp public groups are very similar to social networks. Thus, public groups have been used to spread misinformation, especially as part of articulated political or ideological campaigns [Vosoughi et al. 2018].

In this context, the early detection of misinformation could prevent its spread, thus reducing its damage. This tutorial provides an overview of recent developments in monitoring misinformation spreading, automatic misinformation detection, and identifying misinformation spreaders.

---

[1]DATASENADO. Redes Sociais, Notícias Falsas e Privacidade de Dados na Internet. Brasília, nov. 2019. Available in: www2.camara.leg.br/a-camara/estruturaadm/ouvidoria/dados/pesquisa-nov-2019-relatorio-completo. Accessed in: 26 abr. 2020.

[2]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.

2. Related Work

In [de Sá et al. 2021], we presented the Digital Lighthouse[3], an entire platform for finding, gathering, analyzing, and visualize public groups in WhatsApp. The proposed platform architecture comprises four modules. Module I aims to find WhatsApp public groups. Module II aims to get and store the messages circulating in that groups. Module III explores the data stored in the platform to find implicit, previously unknown, and potentially useful patterns. Module IV explores data visualization concepts to represent information graphically, highlighting patterns and trends in data and helping to achieve new insights.

In [Cabral et al. 2021] we built a large-scale, labeled, anonymized, and public dataset formed by WhatsApp messages in Brazilian Portuguese (PT-BR), collected from public WhatsApp groups [4]. Then, we conduct a series of classification experiments using combinations of Bag-Of-Words features and classical machine learning methods, resulting in a total of 108 combinations, in order to build a specif MID for WhatsApp messages. Our best results achieved a F1-score of 0.733, which may serve as a baseline for future work. The previous results showed that trustful misinformation detection in WhatsApp messages is still a open problem. As a practical result of this work, we built and deployed a Misinformation Detector, which receives a text as input and returns as output the probability that the text contains some misinformation [5].

In [Martins et al. 2021a], we presented a large-scale, labeled, and public data set of WhatsApp messages in Brazilian Portuguese about coronavirus pandemic, called COVID-19.BR. In addition, we performed a wide set of experiments seeking out to build an efficient solution to the MID problem in this specific context. Our best result achieved an F1 score of 0.778 due to the predominance of short texts. However, when texts with less than 50 words are filtered, the F1 score rises to 0.857. All the experiments and the COVID-19.BR data set are available at our public repository[6].

In [Martins et al. 2021c] we proposed a new approach, called MIDeepBR, based on BiLSTM neural networks, pooling operations and attention mechanisms. MIDeepBR can automatically detect misinformation in PT-BR WhatsApp messages. MIDeepBR will automatically detect misinformation at the Digital Lighthouse [de Sá et al. 2021] platform. Experimental results evidence the suitability of the proposed approach to automatic misinformation detection. Our best results achieved an F1 score of 0.834, while in previous works [Martins et al. 2021a], the best results achieved an F1 score of 0.778. Thus, MIDeepBR outperforms our previous work.

In [Martins et al. 2021b] we detail the COVID19.BR data set [Martins et al. 2021a]. With this data set, researchers will be able to build models to perform automatic misinformation detection. Besides, we present a case study exploring data visualization concepts to represent information graphically, highlighting patterns and trends in data and achieving new insights about misinformation phenomenon on WhatsApp.

---

[3]Currently, the platform can be accessed through the link faroldigital.info

[4]This dataset is available on https://github.com/cabrau/FakeWhatsApp.Br

[5]Currently, the Misinformation Detector can be accessed through the link https://faroldigital.info/classifier/misinformation-text

[6]https://gitlab.com/jmmonteiro/misinformation_covid19

3. Open Problems

Despite the efforts made in recent years, many problems remain open. In this section, we will discuss some of them.

Misinformation detection in text contents: The combination of machine learning and natural language processing (NLP) has gotten great results in automatic misinformation detection in WhatsApp messages. However, new NLP strategies and recent deep learning architectures can be explored to improve the MID accuracy.

Misinformation detection in image, audio and video contents: A notable form of abuse in WhatsApp relies on several manipulated images, memes and videos containing all kinds of fake stories. On the other hand, it has been reported that the use of voice messages on WhatsApp has rapidly increased recently: over 200 million voice messages are sent by WhatsApp app every day in some regions. In this context, it is very important to develop MID approaches to images, videos and audios contents, besides multimodal approaches.

Identification of misinformation spreaders: In general, the Pareto rule also applies in the context of WhatsApp public groups. Thus, about 20% of users are responsible for the circulation of 80% of the misinformation that circulates on this platform. So, identifying these super spreaders users is an extremely relevant task. Recently, some approaches have been proposed with the aim of identifying these users.

Developing games to teach media literacy: Using the datasets built previously we can develop a game to teach people fact-checking skills and how to spot misinformation. There are some interesting games for English, such as: Factitious, Get Bad News, Cast Your Vote and Troll Factory. However, there are few games for Portuguese.

Investigation of multidisciplinary approaches: Misinformation is a multidisciplinary phenomenon, involving not only computing science, but also journalism, sociology, education and law. Each one of these sciences has studied this phenomenon using their own tools and perspectives. In this sense, different proposal have emerged to face the misinformation, such as: specific legislation, media education, economic boycott, regulation of social media platforms and public policies.

A knowledge graph on misinformation in WhatsApp messages: As far as we know, no reasonably large and up-to-date knowledge graph (KG) of structured information about the misinformation circulating in WhatsApp has been made publicly available. The use of a KG can facilitate the execution of structured queries involving both the message texts and their metadata (day of publication, time, state or country of the user, etc.). A KG will allow advanced exploration, for example, through queries such as "find all messages that contain misinformation and mention the STF published in August 2021" or "find the top 5 politicians mentioned in messages that contain misinformation".

A proactive chatbot for misinformation detection: A proactive chatbot can automatically monitor, detect and alert the presence of misinformation. Initially, the chatbot needs to be added to a certain group. Then it will automatically monitor and analyze the content that travels in the group. Finally, if it detects that a certain content has a high probability of containing misinformation, an alert message is sent to the group.

4. Conclusion

This tutorial provides an overview of recent developments in monitoring misinformation spreading, automatic misinformation detection, and identifying misinformation spreaders. In addition, we provide an overview of the leading open problems associated with the misinformation phenomenon and briefly examine some of the existing solutions. We hope that our tutorial can help researchers better understand Brazil's misinformation propagation and use data science methods to face this critical phenomenon.

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 Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *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 Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*, pages 297–304. SCITEPRESS.

Guo, B., Ding, Y., Yao, L., Liang, Y., and Yu, Z. (2019). The future of misinformation detection: New perspectives and trends. *CoRR*, abs/1909.03654.

Martins, A. D. F., da Cunha, L. C. C., Monteiro, J. M., and de Castro Machado, J. (2021a). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *Proceedings of the 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021*. Springer.

Martins, A. D. F., da Cunha, L. C. C., Mourão, P. J. C., de Sá, I. C., Monteiro, J. M., and de Castro Machado, J. (2021b). Covid19.br: A dataset of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *III Dataset Showcase Workshop, DSW 2021, Rio de Janeiro, RJ, Brazil, October 4-8, 2021 (To appear)*. SBC.

Martins, A. D. F., da Cunha, L. C. C., Mourão, P. J. C., Monteiro, J. M., and de Castro Machado, J. (2021c). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages using deep learning. In *XXXVI Simpósio Brasileiro de Banco de Dados, SBBD 2021, Rio de Janeiro, RJ, Brazil, October 4-8, 2021 (To appear)*. SBC.

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.