Towards Automatic Fake News Detection in Digital Platforms: Properties, Limitations, and Applications*

Julio C. S. Reis, Fabrício Benevenuto

[1]Department of Computer Science – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{julio.reis,fabricio}@dcc.ufmg.br

Abstract. *Digital platforms, including social media systems and messaging applications, have become a place for campaigns of misinformation that affect the credibility of the entire news ecosystem. The emergence of fake news in these environments has quickly evolved into a worldwide phenomenon, where the lack of scalable fact-checking strategies is especially worrisome. In this context, this thesis aim at investigating practical approaches for the automatic detection of fake news disseminated in digital platforms. Particularly, we explore new datasets and features for fake news detection to assess the prediction performance of current supervised machine learning approaches. We also propose an unbiased framework for quantifying the informativeness of features for fake news detection, and present an explanation of factors contributing to model decisions considering data from different scenarios. Finally, we propose and implement a new mechanism that accounts for the potential occurrence of fake news within the data, significantly reducing the number of content pieces journalists and fact-checkers have to go through before finding a fake story.*

1. Introduction

1.1. Motivation

Digital platforms, including social media systems and messaging applications, are actively used by over one-third of the world's population[1]. These platforms have significantly changed the way users interact and communicate online, opening a whole new wave of applications, and modifying existing information ecosystems. Particularly, digital platforms have dramatically changed the way news is produced, disseminated, and consumed, opening unforeseen opportunities, and also creating complex challenges.

Part of the reasons for this change are inherent to the nature of these digital platforms: (i) it is often more timely and less expensive to produce and consume news on digital platforms compared with traditional news media, such as newspapers or television; and (ii) it is easier to share, comment on, and discuss the news with friends or other readers in digital platforms, which enhances communication and interactions among users. Hence, digital platforms are shaping the way users consume information. Nowadays, about 62% of US users and 66% of Brazilian users get news from digital platforms[2][3]. Despite the

*This work relates to a Ph.D. thesis defended in the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais on November 3, 2020. Authors order: Ph.D. Candidate and Advisor.

[1]http://www.businessinsider.com/the-social-media-demographics-report-2017-8

[2]http://www.pewresearch.org/fact-tank/2016/07/07/modern-news-consumer/

[3]https://reutersinstitute.politics.ox.ac.uk/risj-review/statistic-week-how-brazilian-voters-get-their-news

numerous benefits that these systems bring to our society, they have become a place for campaigns of misinformation which are often intended to deceive people, especially in contexts such as health and politics.

Regarding health, the flood of fake medical news disseminated on digital platforms is causing irreparable damage [Dai et al. 2020]. For instance, a cancer patient mistook an online ad for experimental cancer treatment as medically reliable information, which resulted in his death[4]. Furthermore, during the COVID-19 pandemic, there has been an uptick in rumours and conspiracies spreading through social platforms [Ferrara 2020]. The International Fact-Checking Network (IFCN)[5] found more than 3,500 false claims related to COVID-19 in less than two months[6]. As result, at least 800 people may have died around the world because of coronavirus-related misinformation in the first three months of 2020[7].

In the political context, election after election, we can see different forms of misconduct and complex strategies of opinion manipulation through the spread of fake news. The 2016 presidential election in the USA is still remembered for a 'misinformation war' that happened mostly through Twitter and Facebook. The notorious case involved an attempt of influence from Russia through targeting advertising[8]. Similar attempts were observed during the 2018 Brazilian elections, where WhatsApp was abused to send out misinformation campaigns, with large use of manipulated images and memes containing all kinds of political attacks. A recent study showed that 88% of the most popular images shared in the last month before the Brazilian elections were fake or misleading[9]. Also using WhatsApp, in India, fake rumors spread through the online service were responsible for multiple cases of lynching and social unrest [Arun 2019].

A unique characteristic of news in digital platforms that supports this phenomenon of fake news is that anyone can register/behave as a news publisher without any upfront cost (e.g., anyone can create a Facebook page claiming to be a newspaper or news media organization, or yet, create a group on WhatsApp to spread news). Consequently, not only traditional news corporations are increasingly migrating to digital platforms, but also many news outlets are also emerging on these environments. For instance, previous efforts showed that in 2018 there were more than 20 thousand pages in the USA categorized as news publishers on Facebook [Ribeiro et al. 2018], and this number is continuously growing.

Along with this transition, there are growing concerns about fake news publishers producing and posting fake news stories, and often disseminating them widely through digital platforms [Lazer et al. 2018]. For instance, a study funded by Avaaz[10] asked Brazilian voters whether they saw and believed in five of the most popular fake news on digital platforms during the last weeks of the election in 2018. Impressively, the results revealed that over 98% of interviewed voters were exposed to one or more fake news ar-

[4]https://www.bbc.com/news/business-36189252

[5]The IFCN (https://www.poynter.org/ifcn/) is a unit of the Poynter Institute dedicated to bringing together fact-checkers worldwide.

[6]https://www.poynter.org/coronavirusfactsalliance/

[7]https://www.bbc.com/news/world-53755067

[8]https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html

[9]https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html

[10]www.avaaz.org/

ticles and almost 90% of them believed that these stories were true[11]. Potentially, these numbers impacted the democracy in Brazil face of the 2018 presidential elections.

Misinformation, spin, lies, and deceit have been around forever, but the rise of digital platforms has potentially increased the spread of misinformation and thus have turned the problem of fake news into a worldwide phenomenon, where the lack of scalable fact-checking strategies is especially worrisome. Therefore, in this scenario, our hypothesis is that automatic fake news detection can have a useful degree of discriminative power to identify content that is more likely to be fake, supporting the fact-checking process as well as minimizing the impact caused by extensive production, dissemination, and consumption of fake news through digital platforms.

1.2. Goals

In this thesis, we investigate the potential of automatic solutions to identify fake news disseminated on digital platforms. Whereas fact-checking is an essential strategy to identify fake news that is simple but does not scale, automatic solutions for fake news detection could be used as an assistive tool for fact-checkers to identify content that is more likely to be fake or content that is worth checking, still leaving the final call to an expert at the endpoint of the process. Furthermore, these strategies could be incorporated by digital platforms and search engines as a way to limit the audience of suspicious news stories.

However, automatically identifying fake news is not a trivial task. First, humans themselves are naturally limited at differentiating between real and fake news [Shu et al. 2017], especially when it comes to sensitive subjects, such as politics and health. In addition, news stories are produced by different sources in which each one has its own content style and intrinsic bias, and they are disseminated in different ways through distinct environments, which makes the fake news identification task even harder. Thus, each of these aspects of news (i.e., content, source, environment) can be modeled according to a different set of features that can allow an understanding of typical patterns of fake news that hold across different scenarios. Assessing those differences is crucial to enable the development of language/culture agnostic models for fake news detection.

Therefore, we explore features and solutions that remain useful considering different scenarios and investigate strategies with practical potential for detecting fake news spread on digital platforms.

2. Contributions and Results

2.1. Assessing the Prediction Performance of Solutions to Detect Fake News

There are some current research efforts aiming to understand fake news phenomena and to identify typical patterns and features for proposing automatic solutions for fake news detection. Despite the undeniable importance of the existing efforts in this direction, they are mostly concurrent work which identifies recurrent patterns on fake news after they have been already disseminated or that propose new features for training classifiers using data from a specific scenario, based on ideas that have not been tested in combination. Thus, it is difficult to gauge the practical potential that supervised models trained from

[11]https://www1.folha.uol.com.br/poder/2018/11/90-dos-eleitores-de-bolsonaro-acreditaram-em-fake-news-diz-estudo.shtml

features proposed in recent studies have for detecting fake news. Thus, we first survey a large number of recent and related works as an attempt to implement all potential features to detect fake news. In this field, we explore data from different scenarios: the 2016 US Election and Health. We also build a new dataset containing news stories disseminated through images on WhatsApp during the 2018 Brazilian presidential election. Moreover, we also propose 22 new features for fake news detection, such as those related to the news source (e.g., domain, information regarding source credibility), an indicator of toxicity of text, image safe search indicators, external propagation measure (outside digital platforms) and readability features to assess the writing style of news stories, which have a useful degree of discriminative power for detecting fake news. We then evaluate and compare different supervised machine learning approaches, assessing their prediction performance in the task of automatically identify fake news disseminated in different scenarios. Our results reveal that automatic fake news detection could be used by fact-checkers as an auxiliary tool for identifying content that is more likely to be fake.

2.2. Quantifying the Informativeness of Features for Fake News Detection

Another open issue is that little is known about the discriminating power of features proposed in the literature for fake news detection, either individually or when combined with others, especially involving different scenarios. Some may be adequate for pinpointing fake news with specific patterns, while others are more general but not sufficiently discriminating. Moreover, while explaining the decisions made by the proposed algorithms for fake news detection is central to understand the structure of fake content, this discussion is often left aside. We address all these issues in this thesis. Specifically, after assessing the prediction performance of current supervised machine learning approaches and features for automatic detection of fake news, we provide answers to the following questions. *Do we need all proposed features for fake news detection, or should we focus on a smaller set of more representative features? Is there a trade-off between feature discriminating power and robustness to pattern variations? Is there a clear link between features and the patterns of fake news they can detect?* Since the considered features for fake news detection may have a variety of complex nonlinear interactions, we propose a framework for quantifying their informativeness. In addition, we build models employing a fast and effective classification algorithm with significant flexibility and propose an unbiased strategy to generating them, which enables to perform a unique macro-to-micro investigation of the considered features. We hypothesize that there is no single model to tackle all facets of fake news detection, suggesting that understanding the informativeness of specific combinations of features can be useful for building robust models capable of identifying fake news with different patterns. To accomplish this task, we explore the data from different scenarios as previously mentioned. As part of our proposed framework, we also present an explanation of factors contributing to model decisions, thus promoting civic reasoning by complementing our ability to evaluate digital content and reach warranted conclusions. Last, we investigate whether *there is a set of features that yield models with high performance and able to identify fake news disseminated on digital platforms considering data from different scenarios*, i.e., the 2016 US and 2018 Brazilian presidential elections. In order to accomplish such a goal, we propose an experiment based on Pareto-Efficiency, which is a central concept in Economics widely explored in several areas of knowledge, including the Computer Science. Our findings reveal that fake news with different patterns tend to be identified by models with specific combinations of features. As

result, different models separate fake stories from real/unchecked ones based on very different reasoning. This shows the complexity of the problem and allows us to understand how hard it is for a single solution to tackle all forms of fake news stories.

2.3. Exploring the Practical Potential of Fake News Detection

Last, we explore our findings towards automatic fake news detection to develop a new strategy to help fact-checkers identify news stories that are more likely to be fake, incorporating our approach into a real system called the WhatsApp Monitor (`http://www.whatsapp-monitor.dcc.ufmg.br/`), developed as part of the "Eleições Sem Fake" project (`http://www.eleicoes-sem-fake.dcc.ufmg.br/`). It is a web-based system that helps researchers and journalists by ranking content shared on WhatsApp public groups and displaying them in an organized way. This tool has been used by many journalists and agencies, including Comprova, a collaborative journalistic project from First Draft focused on verifying questionable stories published on social media and WhatsApp during the 12 weeks leading up to the Brazilian 2018 presidential election. However, it only displays a list of the most frequently shared content in the monitored groups over a time interval. This does not necessarily indicate which content should be fact-checked first, as the popularity of a news story in WhatsApp may not be representative of its popularity elsewhere. Therefore, we propose and implement a new mechanism that accounts for the potential occurrence of fake news within collected data, significantly reducing the number of content pieces journalists and fact-checkers have to go through before finding a fake story. Specifically, we explore the machine learning methods to estimate a *fakeness score* on news stories aiming at improving ranking results, which can support decisions regarding the selection of facts (or news) to be checked. Last, we deploy our approach in the WhatsApp Monitor. Our experimental evaluation shows that this tool can reduce by up to 40% the amount of effort required to identify 80% of the fake news in the data when compared to current mechanisms practiced by the fact-checking agencies for the selection of material to be checked such as popularity ranking.

3. Academical and Social Impacts

The main results of this thesis appear in the following publications: [Reis et al. 2020b, Reis et al. 2019b, Reis et al. 2019a, Reis et al. 2017] and [Reis et al. 2016] best paper honorable mention. Further, this thesis opens a novel dataset to the research community containing fact-checked fake images shared through WhatsApp during the 2018 Brazilian presidential election [Reis et al. 2020a], as aforementioned. This dataset can be found in the following link: `http://doi.org/10.5281/zenodo.3779157`.

It is worth mentioning that our work has been awarded by 2018 *Google Research Awards for Latin America* and nominated by Brazilian Symposium on Information Systems (CTDSI/SBSI2021[12]) among the best thesis to compete in the national award. Furthermore, during the development of this thesis, we were also involved in related studies, and embraced many opportunities to collaborate with other researchers. More importantly, all these efforts have received recognition from the academic community and some of them are becoming widely cited[13].

[12]https://sbsi2021.facom.ufu.br/pages/pt/artigos.html

[13]https://scholar.google.com/citations?hl=pt-BRuser=NLkndmAAAAAJ

In terms of social impact, the results obtained from this thesis inspired the construction of systems that were incorporated into the Eleições Sem Fake project, which was extensively used during the Brazilian 2018 elections. Moreover, the Ministério Público de Minas Gerais (MPMG) is supporting a project, two years (2021-2022), in which the goal is to deploy the framework for automatic fake news detection proposed in this thesis integrated with tools for monitoring information disseminated on digital platforms.

Acknowledgements. This work was partially supported by MPMG, project Analytical Capabilities, as well as grants from Google, CNPq, CAPES, and Fapemig.

Referências

Arun, C. (2019). On whatsapp, rumours, and lynchings. *Economic & Political Weekly*, 54(6):30–35.

Dai, E., Sun, Y., and Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 853–862.

Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, 25(6).

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Reis, J., Miranda, M., Bastos, L., Prates, R., and Benevenuto, F. (2016). Uma análise do impacto do anonimato em comentários de notícias online. In *Proc. of the Brazilian Symposium on Collaborative Systems (SBSC)*, pages 46–60.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019a). Explainable machine learning for fake news detection. In *Proc. of the Int'l ACM Conference on Web Science (WebSci)*, pages 17–26.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019b). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

Reis, J. C., Kwak, H., An, J., Messias, J., and Benevenuto, F. (2017). Demographics of news sharing in the us twittersphere. In *Proc. of the ACM Conference on Hypertext and Social Media (HYPERTEXT)*, pages 195–204.

Reis, J. C., Melo, P., Garimella, K., Almeida, J. M., Eckles, D., and Benevenuto, F. (2020a). A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 903–908.

Reis, J. C., Melo, P., Garimella, K., and Benevenuto, F. (2020b). Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review*.

Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 290–299.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorat. Newsletter*, 19(1):22–36.