A Comprehensive Dataset of Brazilian Fact-Checking Stories

Igor Marques[1], Isadora Salles[1], João M. M. Couto[1], Breno C. Pimenta[1], Samuel Assis[1],
Julio C. S. Reis[2], Ana Paula C. da Silva[1], Jussara M. Almeida[1], Fabrício Benevenuto[1]

[1] Universidade Federal de Minas Gerais, Brazil
[2] Universidade Federal de Viçosa, Brazil
{igorma,isadorasalles,joaocouto,brenopimenta,samuelassis}@dcc.ufmg.br
jreis@ufv.br,{ana.coutosilva,jussara,fabricio}@dcc.ufmg.br

Abstract. In recent years, digital platforms have become a powerful means for large scale information diffusion worldwide, particularly in Brazil. Understanding key aspects driving the misinformation diffusion process is of paramount importance to the design and implementation of new tools to automatically detect misinformation content. In this scenario, fact-checking performed by high credibility agencies provide rich labeled data, which is fundamental to build tools capable of detecting and mitigating the effects of misinformation. This paper opens a novel dataset, referred to as *FactCenter*, to the research community, containing fact-check instances collected from 6 different Brazilian fact-checking agencies. This dataset has 11 647 fact-check instances, covering several topics and domains. We present an initial analysis of the data collected, enriched by data from Facebook, which demonstrates the potential of our repository for future studies.

Categories and Subject Descriptors: I.2 [**Artificial Intelligence**]: Natural Language Processing; I.7 [**Document and Text Processing**]: Miscellaneous; J.4 [**Computer Applications**]: Social and Behavioral Sciences; K.4 [**Computing Milieux**]: Computers and Society

Keywords: Fact-checking, Social Media, Misinformation, Fake news

1. INTRODUCTION

In recent years, digital platforms have become a powerful means for large scale information diffusion worldwide, and in Brazil, in particular Newman et al. (2019). Despite offering an unprecedentedly effective method of spreading relevant information, from entertainment to politics to large populations, digital platforms have also been reportedly exploited to host campaigns of misinformation dissemination Tardaguila et al. (2018); Resende et al. (2019). Such practice has been employed with various controversial purposes, from public opinion manipulation in the context of democratic processes Bessi and Ferrara (2016); Gomes Jr and Frizzon (2019); Machado et al. (2019) to the encouragement of questionable health behaviours Ferrara (2020); Martins et al. (2021). In all scenarios, it has the potential to boost radicalization and intensify social conflicts even beyond the online world Ribeiro et al. (2019).

A plethora of research efforts and initiatives have been emerging to fight and mitigate the impact of online misinformation spread Myslinski (2012); Vlachos and Riedel (2014a); Wu et al. (2014); Reis et al. (2019); Reis and Benevenuto (2021). One of the most known initiatives is the emergence of fact-checking agencies, which have the goal of verifying the level of truthfulness of the information disseminated over digital platforms. Examples of international fact-checking agencies are *Snopes.com*[1],

[1] www.snopes.com

Copyright©2022 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

PolitiFact[2], *FactCheck.org*[3]. In Brazil, widely known fact-checking agencies include *Aos fatos*[4], *Agência Lupa*[5], *Boatos.org*[6], *Comprova*[7]. The challenges faced by these agencies are (at least) twofold. First, misinformation content is disseminated on different platforms and has various formats, such as news, videos, memes, and audio, which ultimately requires additional effort. Thus, it is difficult to track misinformation content through a standard procedure. Second, low credibility publications are produced at a fast pace, often at a much higher rate than the agencies are able to fact-check them Ciampaglia et al. (2015).

Understanding key aspects driving the misinformation diffusion process is of paramount importance to the design and implementation of new tools to automatically detect misinformation content. Fact-checking performed by high credibility agencies provide rich labeled data, which is fundamental to build tools to detect and mitigate the effects of misinformation. However, there have been few attempts to unify fact-checking contents focused on specific contexts (e.g. health[8]) and mainly these efforts are only found written in English. Additionally, solutions that focus on building high credibility unified fact-checking repositories, annotated by experts (e.g. journalists) with domain experience Reis et al. (2020) and that cover different domains (such as health, politics) and topics (COVID-19, political elections) are essential to leverage studies aimed at understanding and preventing large-scale dissemination of misinformation in Brazil.

In light of the previous discussion, in this work, which builds upon our prior effort Couto et al. (2021), we detail a methodology for gathering and organizing fact-check instances from 6 different Brazilian agencies in a centralized database: *Agência Lupa, Aos Fatos, Boatos.org, Comprova, Estadão Verifica* and *Fato ou Fake*. We refer to our repository as *FactCenter* and in order to allow reproducibility and to foster follow-up studies, we have released it for public use[9]. In total, we gathered 11 647 fact-check instances published between July 2013 and May 2021, covering several topics, such as the COVID-19 pandemic and several Presidential elections, as well as various domains, such as health and politics. Compared with our prior work Couto et al. (2021), here we offer a deeper analysis of this dataset. First, we perform a series of general analysis, including the volume of checks performed per agency over time, to offer an overview of our data. We then focus on textual analysis, extracting and analyzing topics from the fact-check instances. We also enrich our analysis by crawling data from the agencies' Facebook pages, through the Crowdtangle platform[10]. This data allows us to study the agencies' popularity as well as the people's reactions towards the information fact-checked by them. More importantly, it sheds light on how this type of content attracts attention on online social networks, helping on the proposal of mechanisms to fight misinformation in these environments.

Our analyses unveiled a number of interesting findings. We found that there is a trend towards an increase in the number of fact-check instances, possibly due to the reportedly increasing spread of misinformation on different digital platforms. Moreover, the most frequent topics in our dataset are characterized by words such as *social, fake, health, misinformation* and *Bolsonaro* (Brazilian president). We also found that fact-checking organizations attract a large number of followers on Facebook with a non-negligible level of interaction through reactions.

The rest of this paper is organized as follows. Section 2 summarizes prior related work. Section 3 describes our methodology to gather *FactCenter* dataset, including an overview of it, while Section 4 presents the analyzes performed in this work. Finally, Section 5 concludes the paper, presenting some

[2] www.politifact.com/
[3] www.factcheck.org/
[4] aosfatos.org
[5] piaui.folha.uol.com.br/lupa/
[6] www.boatos.org
[7] projetocomprova.com.br/
[8] https://www.poynter.org/coronavirusfactsalliance/
[9] https://doi.org/10.5281/zenodo.5191798
[10] https://www.crowdtangle.com/

potential research directions that might benefit from the data we made publicly available.

2. RELATED WORK

A large body of recent studies investigated the phenomenon of misinformation in digital platforms (i.e. social media systems, messaging applications, etc). These studies can be roughly divided into two main groups: (i) efforts aiming at understanding the phenomenon per se or proposing solutions to detect and mitigate it's effects and; (ii) efforts to gather data to support the outcomes of (i), providing publicly available labeled fact-check datasets. The following paragraphs summarize some of these efforts.

An exemplar work of category (i), Vosoughi et al. (2018) analyze the spread of news stories true and false (as verified) on Twitter from 2006 to 2017, showing that false news spread significantly farther, faster, deeper, and more broadly than truthful ones. Lazer et al. (2018) discuss social and computer science research regarding the belief in fake news and the mechanisms by which it spreads. Resende et al. (2019) analyze the messages shared on a number of political oriented WhatsApp groups, focusing on textual content. Using a dataset of fact-checked misinformation instances from six Brazilian fact-checking sites, they identify the presence of misinformation in the contents of these messages . Authors in (Resende et al., 2019) find that images are the most popular type of media content shared on publicly accessible WhatsApp groups related to politics during two major political events in Brazil. They propose a methodology to automatically identify images containing misinformation, and used it to investigate the sharing of this type of content in the monitored groups. Further, Maros et al. (2021) analyze more than forty thousand audio messages shared in over 364 publicly accessible WhatsApp groups in Brazil. They focus their analyses on content and propagation properties of misinformation found in audio messages, contrasting them with unchecked content as well as with prior findings about misinformation in other media types. Authors find that audio messages with misinformation tend to spread quicker than unchecked content and remain active significantly longer in the network.

Beyond characterization studies, there have been efforts to use machine learning methods to automatically detect misinformation. Castillo et al. (2011) analyze the credibility of news propagated through Twitter and proposed a classifier to automatically determine which topics are newsworthy, assigning to each newsworthy topic a credibility label. Volkova et al. (2017) propose linguistically-infused neural network models to classify social media posts retweeted from news accounts into verified and suspicious categories – propaganda, hoax, satire and clickbait. Shu et al. (2017) offer an extensive review of existing literature on misinformation detection approaches from a data mining perspective, including techniques for feature extraction and model construction. Reis et al. (2019) survey existing studies on applying supervised learning models to detect fake news, identifying the main features in use. The authors implement those features and test their effectiveness via a variety of supervised learning classifiers trained to distinguish fake from real stories on a large labeled dataset. Córdova Sáenz et al. (2021) propose a fake news classification process based on the combination of textual content of the news and the topology of the news diffusion networks. Authors propose the use of DistilBERT Sanh et al. (2020) to generate features that compactly characterized the news. To account for the social context of each news instance, they propose to represent the properties of the diffusion network of each news item on Twitter by topological metrics, including tweets, retweets and mentions.

Towards designing solutions to fight misinformation, researchers need a broad set of datasets containing labeled data, i.e., fact-check content, covering different topics and contexts. Most of the publicly available datasets (works in (ii)) contain English-language content (Reis et al., 2020). As examples, *LIAR* (Wang, 2017) is a dataset of short statements from PolitiFact.com manually labeled as half-true (2 638), false, mostly-true, barely-true, true and pants-fire. *BuzzFace* (Potthast et al., 2018) gathers news published on Facebook from 9 agencies over a week close to the 2016 U.S. election. *Fact-Checked-Stat* (Vlachos and Riedel, 2014b) is a dataset containing a list of statements fact-checked from popular fact-checking websites labeled by journalists. *Fake-News-Net* (Shu et al., 2017) is a repository for an ongoing data collection project for fake news research. *Fake-Satire* (Golbeck et al., 2018) is a dataset of fake news and satire that are hand-coded and verified. Lazer et al. (2018) introduce a fake news dataset around the Syrian war (FA-KES). *Fake-Twitter-Science* (Vosoughi et al., 2018) is composed of verified true and false news distributed on Twitter from 2006 to 2017. The data comprises around 126 000 instances (rumors cascades) tweeted by 3 million people more than 4.5 million times. Finally, Kaggle[11] is a repository of text and metadata from fake and biased news sources found on the web.

In the context of Brazil, a very limited number of open datasets are available. Specifically, the Kaggle platform offers a dataset of rumors checked by the *Boatos.org* Brazilian agency. This dataset contains 1 900 instances of rumors and the respective fact-check verdicts. Moreno and Bressan (2019) present *FACTCK.BR*, a dataset useful to study fake news instances written in Portuguese, which contains 1 309 fact-check instances from 3 Brazilian agencies: *Aos Fatos*, *Agência Lupa* and *Truco*. The data was collected by the ClaimReview[12] Project. Last, Reis et al. (2020) present a dataset of fact-check images shared in WhatsApp during the Brazilian and Indian Elections.

In an attempt at contributing to such a small number of public collections of misinformation in Portuguese, our work provides a large and rich dataset of Brazilian fact-checks, the *FactCenter* dataset, unique for offering a large total number of fact-check instances (11 647) and for the diversity of topics and domains covered by the dataset. Moreover, our work presents an initial characterization of the data presented, enriched by the Facebook data we collected via CrowdTangle. These analyses offer insights into the popularity of fact-checking in social media as well as on how misinformation may spread in these platforms.

3. DATASETS COLLECTION AND OVERVIEW

In this section, we first present the methodology to build the *FactCenter* dataset (Section 3.1), and then we offer an overview of the collected data (Section 3.2). In Section 3.3, we briefly describe the Facebook data as used to enrich our analyses of the agencies as presented in Section 4.

3.1 *FactCenter* Dataset Construction

As previously mentioned, our main contribution is the construction of a high-quality fact-checking dataset from a set of 6 relevant Brazilian fact-checking organizations. The goal is to gather data labeled by domain experts, from diverse domains and topics, which may be applied to better understand the intrinsic characteristics of misinformation in Brazil.

Figure 1 shows the procedure utilized to build our dataset, henceforth referred to as *FactCenter*. As a first step of our data collection, we selected the fact-checking agencies (step 1). We chose 3 agencies that are verified by the International Fact-Checking Network (IFCN)[13], which establishes the global gold standard that should be observed by fact-checking organizations: *Aos Fatos*[14], *Estadão Verifica*[15] and *Agência Lupa*[16]. To complement our initial selection, we also collected data from three other nationally recognized fact-checking agencies, namely: *Boatos.org*[17], *Comprova*[18] and *Fato ou Fake*[19]. Among those selected the oldest agency was founded in 2013 (*Boatos.org*) and the newest ones in 2018 (*Estadão Verifica* and *Fato ou Fake*).

[11] `https://www.kaggle.com/rogeriochaves/boatos-de-whatsapp-boatosorg`

[12] `https://www.claimreviewproject.com/`

[13] `https://www.poynter.org/ifcn/`

[14] `https://www.aosfatos.org/`

[15] `https://politica.estadao.com.br/blogs/estadao-verifica/`

[16] `https://piaui.folha.uol.com.br/lupa/`

[17] `https://www.boatos.org/`

[18] `https://projetocomprova.com.br/`

[19] `https://g1.globo.com/fato-ou-fake/`

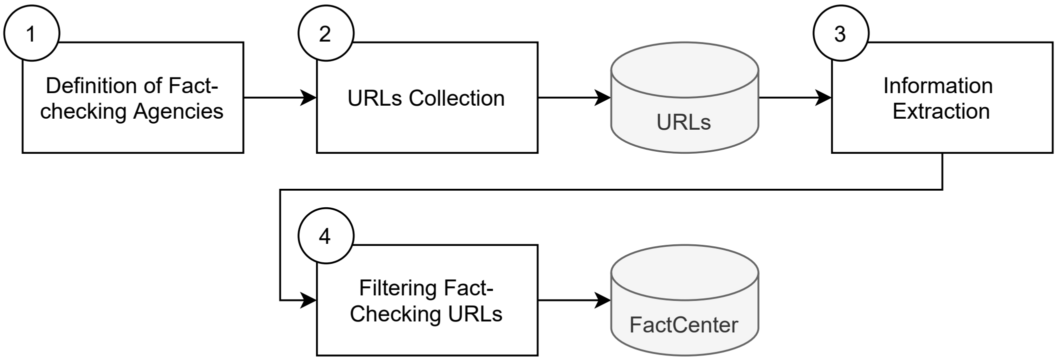

1
Definition of Fact-checking Agencies
2
URLs Collection
URLs
3
Information Extraction
4
Filtering Fact-Checking URLs
FactCenter

Fig. 1: Dataset creation overview.

We then scraped the official website of each agency (step 2) collecting all the URLs available. It is important to note that these websites are composed not only of fact-checking instances, but also news, articles, and other forms of content. Thus we developed scrapers that go through the history of each site in order to access the largest amount of content available. Notably, initially all links found were collected, regardless of whether or not they were associated with a fact-check. The process was performed on all fact-checking websites by making the scraper go through the full extension of the homepage or accessing a section of historical links on the page. At the end, we were able to gather 14 913 URLs, which comprise both fact-check instances and general URLs (e.g., news, advertisements, etc).

Afterwards, using the list of URLs gathered in step 2, we extracted information from each collected URL (step 3). This proved not to be a trivial task provided that pages from fact-checking agencies do not share a clear standardized structure of published content. Thus, we implemented specific crawlers for each fact-checking agency, using Python programming language (version 3). Within this task, we sought to delimit patterns for each agency in order to capture the greatest amount of formatting variation with the least possible loss of information throughout the collected URLs. We then extracted the following data from each fact-check URL:

—`url`: URL to the original fact-check instance;
—`source_name`: Fact-checking agency;
—`title`: Fact-check instance's title;
—`subtitle`: Fact-check intance's subtitle;
—`publication_date`: Publication date (YYYY-MM-DD format, where YYYY, MM and DD represent year, month and day, respectively);
—`text_news`: Fact-check instance's text body;
—`image_link`: URL to image (if available);
—`video_link`: URL to video (if available);
—`authors`: Fact-check instance's author list;
—`categories`: Categories associated with the fact-check instance by the agency;
—`tags`: Keywords associated with the fact-check instance by the agency;
—`obtained_at`: Date that data was collected (YYYY-MM-DD format);
—`verdict label`: Each agency has its own set of labels, for instance true, false, fake, out of context, rumor, etc.

In the following step, we employed filters to identify the fact-check instances and avoid incomplete information on the data collected (step 4). The element utilized to distinguish fact-check URLs from those that are not is the presence of a *verdict label.* The first filter employed at step 4 discards the URLs in which the *verdict label* is absent. Then, a second filter selects the URLs in which we were able to collect at least the *title*, *publication date* and *text* fields. Those fields contain information needed to perform various different types of analysis, such as the ones we propose in Section 4. The links that met the above criteria constituted the dataset with the data fields extracted from them. Therefore, after filtering in the fact-check instances and preventing incomplete data collections, refered to as *fact-checking filter* in Figure 1 (step 4), we discarded a total of 3 266 URLs, which were not identified as containing fact-checks or all essential data fields. At the end of this process, our dataset contains a total of 11 647 instances, published between 2013 and 2021, available in the CSV format at `https://doi.org/10.5281/zenodo.5191798`.

Finally, there are some particularities about *FactCenter* that are worth mentioning. First, fact-checking agencies may provide multiple labels for a single fact-check instance. For instance, it is common for claims in political speeches to reference multiple distinct topics, thus, each topic may have a different verdict associated with it. For these cases, our dataset provides a list of all verdicts assigned to each fact-check instance. Second, some verdicts are not provided in textual format, but through images instead. In those cases, we employed *Optical Character Recognition (OCR)* techniques to extract the textual verdict from the images. Third, potential misinformation instances are often checked by multiple agencies. That being the case, we decided to retain those duplicates in *FactCenter*, as they may offer different perspectives and conflicting verdicts that may help answering research inquiries. At last, the source that originally released the misinformation instances under analysis in each fact-check instance is not included in the data published by the agencies, and thus is not present in *FactCenter*.

3.2 Overview of *FactCenter*

Table I provides an overview of *FactCenter*, including the time period covered by each agencies' publications, the total number of fact-check instances, the average numbers of words per fact-check and publications released per month as well as the number of fact-check instances that have links to images and/or videos. Overall, the average number of words per instance is similar across all agencies except for Comprova, which provides a significantly higher average. We speculate that this is related to the nature of the fact-check provided by this particular agency: first, each fact-check is collaboratively elaborated by expert journalists from different news outlets; the fact-check is published when at least three of the participating outlets approve the verdict and conclusions reached. This process, known as CrossCheck[20], ends up resulting in a larger average number of words per fact-check. Note that, despite having different active operation periodos, all agencies have a high average number of publications per month. Moreover, a non-negligible number of fact-check instances are linked to external images and videos. For instance, *Boatos.org*, in some of their publications, link to a video presentation of the associated fact-check in which a collaborator alludes to the the original misinformation instance as well as presenting the verdict provided by the agency[21]. Additionally, in some fact-check instances we also found links to the original misinformation images as shared on WhatsApp[22].

Figure 2 offers an overview of the temporal evolution of fact-checking activities by showing the monthly time series of the number of fact-checks produced by each agency. Vertical lines indicate the oldest fact-check by each agency. Despite some fluctuations, there is a general trend towards an increase in the number of fact-check instances, possibly due to the reportedly increase in misinformation spread on different digital platforms. We notice that there are some significant spikes in the volume of fact-check instances, which coincide with relevant events. As an illustration, in October

[20] `https://firstdraftnews.org/about/crosscheck-newsroom/`
`https://projetocomprova.com.br/about/faqs/`

[21] `https://www.youtube.com/embed/-64HM7ifFDI`

[22] `https://glo.bo/3ciEy4U`

Table I: Some statistics of the *FactCenter* dataset.

Agency	First day of collection	Last day of collection	# Fact-check instances	Avg. #words per instance	Avg. #publications per month	# URLs to images	# URLs to videos
Agência Lupa	11-25-2015	04-19-2021	2 574	620	40.46	2 574	-
Aos Fatos	10-09-2015	05-13-2021	1 679	756	25.49	1 206	-
Boatos.org	07-01-2013	05-14-2021	5 523	639	59.59	2 037	1 710
Comprova	07-16-2019	05-19-2021	361	1 813	16.63	98	-
Estadão Verifica	08-09-2018	05-19-2021	593	806	18.13	593	-
Fato ou Fake	07-30-2018	04-21-2021	917	553	28.54	809	42

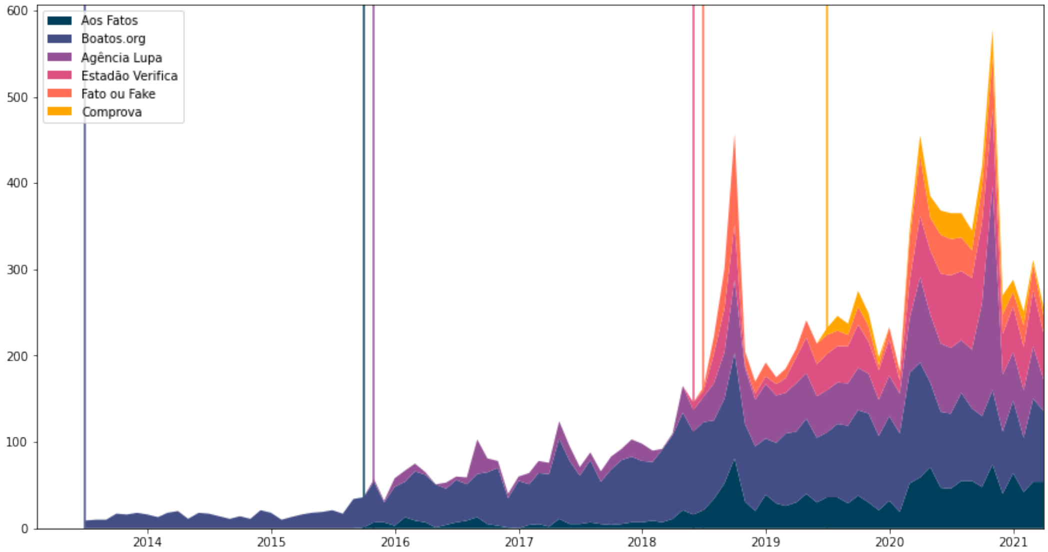


Fig. 2: Temporal evolution of the total number of fact-check instances for each agency. Vertical lines highlight the oldest fact-check of each agency.

2018 during the presidential election in Brazil, the number of fact-checks increased by 152% compared to September 2018. Moreover, we also highlight the huge growth in the number of fact-checks in 2020, probably due to the large-scale spread of misinformation content associated with the COVID-19 pandemic Ferrara (2020).

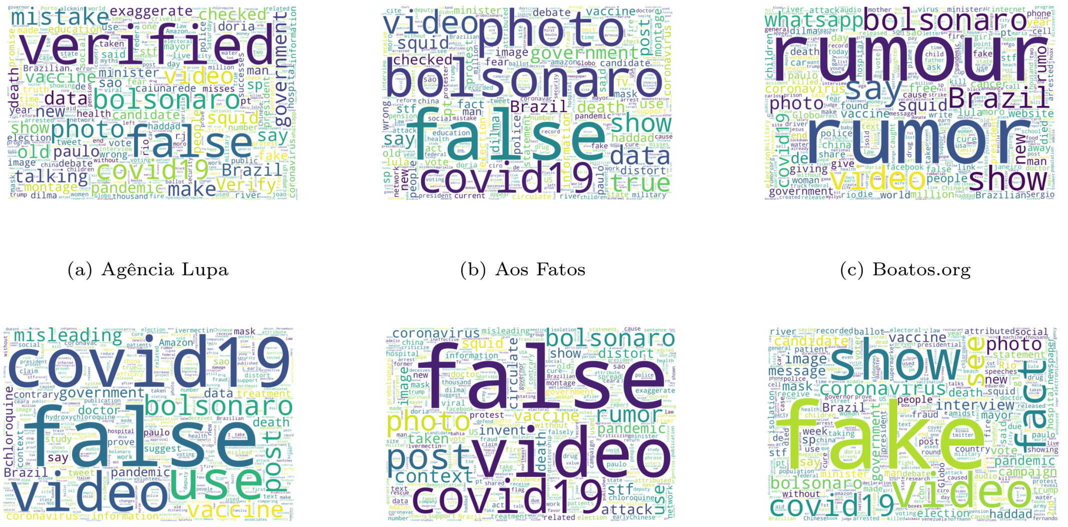
Apart from the major events that may boost the activity of all agencies, the intrinsic characteristics of each agency may also impact the volume of fact-checking produced by them. For instance, fact-check instances from *Boatos.org*, which has the highest volume of instances out of all six agencies, tend to be short and focused on specific topics or claims. *Aos Fatos*, in turn, tends to adopt an opposite strategy, aggregating several checks of a topic in a single text (for instance, different statements made about a particular political party).

FactCenter provides the verdicts as annotated by expert collaborators from the fact-checking agencies. Notably, each agency employs its own set of labels as verdicts, as shown by the label distribution per agency in Table II. Note that *Agência Lupa* and *Aos Fatos* have the most heterogeneous set of labels. Overall, we find that *misleading*, *false*, *rumor*, and *fake* labels are the most frequent labels across all agencies, however no further standardization is observed.

Finally, as a general description of the contents of the fact-check instances in *FactCenter*, we show in Figure 3 the word clouds with the top-100 most popular words (translated to English) present in their publication titles. Overall, *false*, *rumor* and *fake* are the most frequent words, which indicates that agencies include the verdict in the instance's title. Besides that, we also observe that the name of the Brazilian president (*Bolsonaro*) and *covid-19* are frequently used words, reflecting a bias towards political-oriented and coronavirus-related content.

Table II: Labels used by each fact-checking agency as verdicts for fact-check instances.

Agency	Labels (translated to English)
Agência Lupa	false (3209), true (1469), exaggerated (866), true but (723), look at (222), contradictory (189), underestimated (108), too early to make a conclusion (108), unsustainable (97)
Aos Fatos	false (2522), true (637), inaccurate (394), exaggerated (239), distorted (125), unsustainable (120), contradictory (65)
Boatos.org	rumor (5523)
Comprova	misleading (183), false (159), verified (9), verified evidence (6), wrong context (4)
Estadão Verifica	false (299), misleading (160), out of context (134)
Fato-ou-Fake	fake (1098), fact (394), not quite (346)



(a) Agência Lupa (b) Aos Fatos (c) Boatos.org

(d) Comprova (e) Estadão Verifica (f) Fato ou Fake

Fig. 3: Word clouds with the top-100 most popular words in the titles (translated to English).

3.3 Facebook Dataset

To gather further data related to different Brazilian online fact-checking organizations we collected data from Facebook through the Crowdtangle API[23]. Crowdtangle is a simple-to-use tool, which allows to download Facebook pages content with their creation date, posts and their publish date, users' reactions and usage reports (such as likes over time, total number of followers, etc). In total, we retrieved 14 516 posts from the creation date of each page till mid September, 2021. In total, 6 857 out of 14 516 posts mentioned fact-check instances in *FactCenter* dataset[24]. Table III presents an overview of the Facebook dataset. The oldest page is the *Boatos.org* page, created in 2013; the newest pages (2018) are of *Fato ou Fake* and *Comprova*, which may explain the unbalanced number of their posts. *Boatos.org* has four times more posts related to fact-check instances on Facebook (4 633 posts) than *Agência Lupa* (1 001 posts), which is the second in the ranking. Data from *Estadão Verifica* is

[23] `https://www.crowdtangle.com/`
Terms of service does not allow to transfer, sell, disclose, or license any content from or content access through the Services without express written consent of CrowdTangle. `https://www.crowdtangle.com/terms`

[24] To check the presence of a fact-check mention (i.e. fact-check URL) on a Facebook post we used the fields *Link* and *Final Link*, which are specified to identify links in posts gathered by the Crowdtangle API (`https://help.crowdtangle.com/en/articles/3213537-crowdtangle-codebook`)

Table III: Overview of Facebook dataset.

Agency	Creation Date	Last day of collection	#Posts	#Fact-check w/ posts
Agência Lupa	11-24-2015	09-27-2021	3 788	1 001
Aos Fatos	06-22-2015	09-27-2021	2 437	840
Boatos.org	06-14-2013	09-27-2021	6 254	4 633
Comprova	06-17-2018	09-24-2021	750	380
Fato ou Fake	06-28-2018	09-24-2021	1 287	2

Table IV: Number of topics found by the LDA.

Agency	#Topics
Agência Lupa	8
Aos Fatos	10
Boatos.org	22
Comprova	22
Estadão Verifica	2
Fato ou Fake	3

not available, since this agency does not have an official Facebook page for fact-checking publications.

4. ANALYSES

We now offer a deeper analysis of the data collected. Section 4.1 presents the topics covered by the fact-check instances included in *FactCenter*. We then focus our attention on understanding people's interactions with the agencies' content through Facebook pages in Section 4.2.

4.1 Topic Modeling

In this section, we characterize the fact-check instances in terms of the topics they convey. To that end, we employed the Latent Dirichlet Allocation (LDA) Blei et al. (2003), a generative statistical model to automatically infer the topics in a collection of documents. Specifically we used the LDA implementation available in the Gensim Python library[25]. For each agency, we applied LDA in the text content of all its fact-check instances.

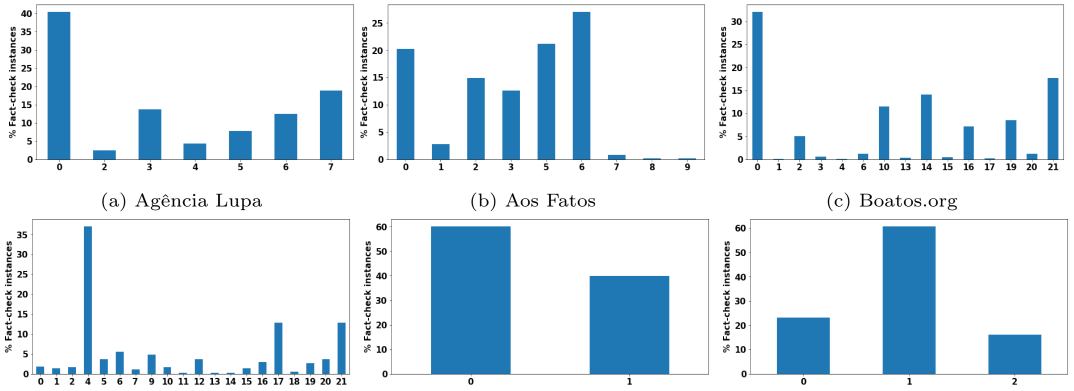
Specifically, we lowercased and tokenized all the words in the text of each fact-check instance, removing numbers, single letters, accents and stopwords using the Portuguese list provided by the NLTK library[26]. For each agency, we ran the LDA algorithm varying the number of topics k from 1 to 25 and choosing the LDA model that produced the highest topic coherence metric Newman et al. (2010). Table IV shows the number of topics selected for each agency. We observe that *Comprova* seems to cover a larger variety of topics, despite the smallest number of fact-check instances from this agency included in *FactCenter*. In contrast, *Estadão Verifica* seems to focus on fewer topics.

For each fact-check instance, we used the derived LDA model from it's respective agency to infer the probability distribution of the topics covered by it's content, and the topic with the highest probability was chosen as the final topic of the instance[27]. Figure 4 shows the histograms of topics covered by the instances for each fact-checking agency. Note that, across all agencies, the most popular topic varies from at least 27% to at most 60% of all fact-check instances from the agency. Table V characterizes the most popular topic for each agency in terms of its most representative words (according to LDA) as well as the number of fact-check instances with the topic (as main topic). We identify that the most frequent topics contain mostly words related to the current Brazilian president (Bolsonaro),

[25] `https://radimrehurek.com/gensim/intro.html`

[26] `https://www.nltk.org/howto/portuguese\_en.html`

[27] In all analyses, each fact-check instance has only one topic assigned to.



(a) Agência Lupa (b) Aos Fatos (c) Boatos.org

(d) Comprova (e) Estadão Verifica (f) Fato ou Fake

Fig. 4: Distributions of topics inferred by LDA.

Table V: Description of topics with the highest number of associated fact-check instances.

Agency	Topic ID	#Fact-Checks	Most relevant words (translated to English)
Agência Lupa	0	1040	facebook, verification, lupa, no, networks, social, project, image, news, circulates
Aos Fatos	6	447	no, facts, networks, also, facebook, fake, social, misinformation, bolsonaro, website
Boatos.org	0	1771	no, internet, history, people, you, boatosorg, whatsapp, already, message, let's
Comprova	4	139	no, covid19, proof, health, also, use, vaccine, treatment, are, coronavirus
Estadão Verifica	0	357	estadao, no, facebook, verify, also, transparency, posts, social, fake, network
Fato ou Fake	1	557	no, message, video, networks, fact, social, team, checks, made, after

Table VI: Popularity statistics based on the Facebook pages' followers.

Agency	Avg. #Followers /#years	Max #Followers	Max #Followers Date	Min #Followers	Min #Followers Date	Current #Followers
Agência Lupa	159 520.8	192 109	07-2021	77 292	07-2017	192 109
Aos Fatos	63 745.2	82 788	07-2021	27 891	07-2017	82 784
Boatos.org	186 494.4	214 694	04-2021	111 930	07-2017	214 534
Comprova	135 045.5	136 751	10-2018	785	06-2018	133 882
Fato ou Fake	122 175.75	154 968	07-2021	3 378	07-2018	154 965

COVID-19 pandemics, social media (Facebook and WhatsApp), media type (image, video, message) and verdict labels (misinformation, fake).

4.2 Agency popularity from the lens of Facebook

Social media is, reportedly, one of the most often used vehicles for misinformation dissemination Newman et al. (2019). Being fact-checking a fundamental strategy to fight misinformation, it is important to understand how people perceive the fact-check instances. In what follows, we use data from Facebook, which is a very popular social media platform, to provide a first look into the popularity of the fact-checking agencies covered in our dataset. Specifically, we rely on the Facebook dataset described in Section 3.3.

We first focus on agency popularity estimated by the total number of followers of each agency's Facebook page. Table VI shows the average number of followers per year, as well as the maximum (Max #Followers) minimum (Min #Followers) and current (at the time of data collection) number of followers, with the date in which maximum and minimum were reached. Recall that *Estadão Verifica* does not have an official Facebook page for fact-checking publications, so this agency is omitted from the table. The two most popular agencies, in average number of followers, *Agência Lupa* and *Boatos.org*, are also those with the largest number of fact-check instances posted in their pages.

Table VII: Total number of reactions to posts and average number of reactions per post.

Agency	#Reactions	#Reactions/#posts
Agência Lupa	1 682 408	444.14
Aos Fatos	380 842	156.27
Boatos.org	1 043 024	166.77
Comprova	1 096 556	1 462.07
Fato ou Fake	364 983	283.59

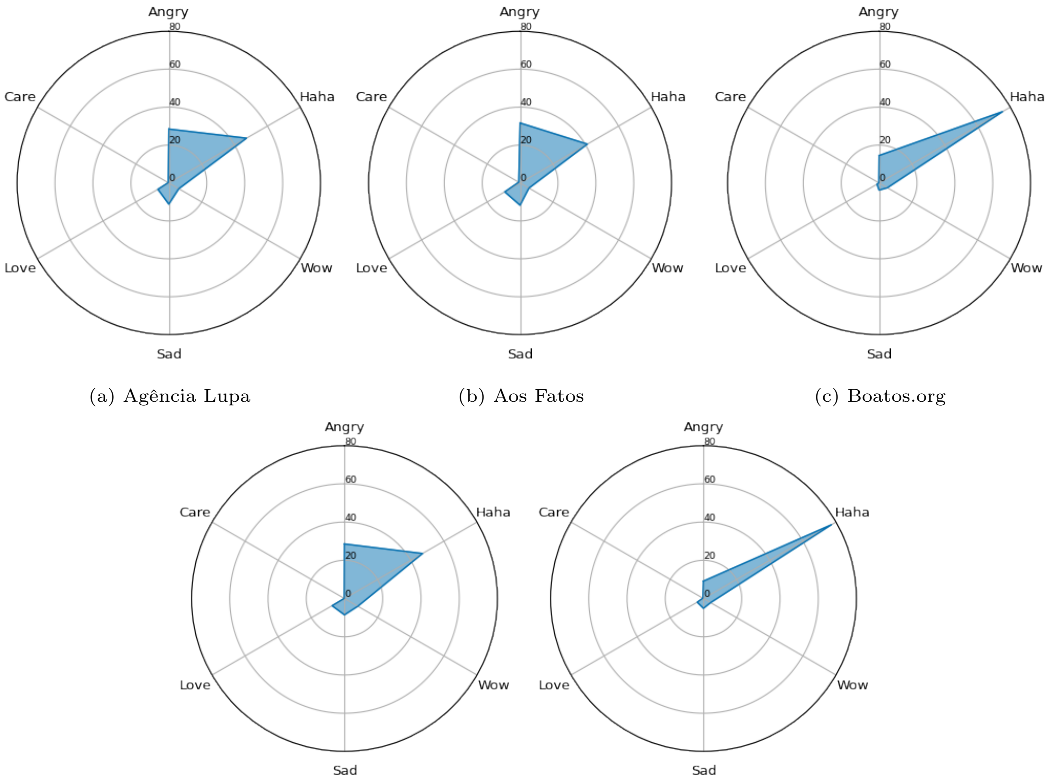
Table VIII: Posts with highest number of reactions.

Agency	Title (Translated to English)	#Likes	#Angry	#Haha	#Wow	#Sad	#Love	#Care
Agência Lupa	#We verified: It is false that Bolsonaro has taken public accounts 'from the red'	3 912	68	644	35	8	126	0
Aos Fatos	How to search information about your candidates before elections	6 419	3	46	4	1	45	5
Boatos.org	7 rumors about Facebook that circulates on the web (and you always buy it)	1 989	8	258	36	1	16	0
Comprova	People who withdraw the R$ 500 from the FGTS will not lose the right of its total balance in case of dismissal; understand the government's proposal	99 453	2 451	4 082	738	286	743	0
Fato ou Fake	It's #FAKE the questioning on police excess in Lázaro's case in Fatima's show	2 337	28	713	12	16	14	2

On Facebook, reactions (*Like, Love, Care, Haha, Wow, Sad and Angry*) are an alternative and quite popular means for people to communicate their feelings towards particular posts in a quick and easy way. As such, the total number of reactions can be a proxy for user engagement towards a particular post, being also a measure of content popularity. Table VII summarizes the total number of reactions attracted by all posts from each agency. On average, followers tend to interact a lot with agencies' posts. For illustration purposes, Table VIII presents, for each agency, the title of the post with the largest total number of reactions, along with the numbers for each reaction type. For the most popular posts, Like is the most prevailing reaction, followed by the Haha reaction.

Towards each agency in general, we broke down the reactions by their type. 82%, 86%, 76%, 85% and 80% of reactions in *Agência Lupa*, *Aos Fatos*, *Boatos.org*, *Comprova* and *Fato ou Fake*, respectively, are Like reactions. Figure 5 depicts the results for all Facebook posts, using radar charts. We omit the Like reaction, which prevails in all agencies, to better visualize the differences among the other reactions in our data. We notice that *Agência Lupa, Aos Fatos* and *Comprova* are very similar, with respect to their reaction popularity distribution, with a balance between the fractions of Haha and Angry reactions. Posts from *Boatos.org* and *Fato ou Fake*, in turn, attract Haha much more often (more than 70%). An interesting future analysis, which is out of the scope of our present work, is to understand if the frequent occurrence of Haha reactions is related to the nature of the fact-check instance, i.e., information highly unlikely to be true, triggering a sarcastic reaction.

We now turn our attention to analyzing how topics with fact-check instances in *FactCenter* attract Facebook followers' attention. To do so, we define a simple user engagement score. Let $\bar{E}_I$ be



(a) Agência Lupa

(b) Aos Fatos

(c) Boatos.org

(d) Comprova

(e) Fato ou Fake

Fig. 5: Posts' reaction popularity distribution, excluding "Like" reactions.

the average of the total number of shares, likes and comments assigned to each fact-check instance mentioned on Facebook. For each topic, we calculated the average engagement score $\bar{E}_T$ over the posts assigned to it. We also calculated $\bar{E}_A$ and $\bar{S}_A$, the agency's average engagement score and standard deviation regardless of the LDA topic. Table IX shows the topics with the highest and lowest values of $\bar{E}_T$ per agency. *Fato ou Fake* is not included once this agency has only two posts related to fact-check instances in *FactCenter*. Note that the topics that engage followers the most are characterized by different words such as *candidate, government, mayor, petista, job, doria*[28], *bolsonaro*[29], *health* and *covid-19*.

5. CONCLUDING DISCUSSION

In this work, we present a novel dataset containing historical data from all fact-check instances published by the key Brazilian fact-checking agencies. We hope this data might be useful for researchers exploring misinformation in Brazil. Next, without any intention of providing an exhaustive list, we describe a few potential directions that could be explored based on this dataset.

Improving the efficacy of fact-checking. All across the globe, dozens of organizations dedicate themselves to verifying the accuracy of claims and stories circulating through our information ecosystem, potentially checking the same story that other organizations have already debunked. It is reasonable to expect that misinformation campaigns reuse or are inspired by conspiracy theories and other misinformation campaigns that were successful in spreading in other countries. Indeed, the

[28] Governor of São Paulo State.
[29] Brazilian President.

Table IX: Description of topics with lowest and highest scores.

Agency	**Topic ID**	$\overline{\mathbf{E}}_\mathbf{A}$	$\mathbf{S_A}$	$\overline{\mathbf{E}}_\mathbf{T}$	**#posts**	**Most relevant words (translated to English)**
Agência Lupa	6	288.20	281.47	479.25	29	no, day, year, million, interview, data, brazil, billion, candidate, government
	4			45.81	7	candidate, october, mayor, city hall, port, november, city, vote, research, agreement
Aos Fatos	1	89.32	210.93	259.36	28	city, no, rio, health, city, data, municipal, are, declaration, mayor
	8			55.83	2	lula, ex-president, senator, trf4, deputy, flavio, petista, lava, mensalao, money
Boatos.org	6	92.86	119.48	200.11	66	covid19, coronavirus, pandemic, virus, health, no, disease, deaths, people, italy
	17			41.67	13	vacancies, job, doria, vacancy, joao, professionals, technician, company, abraham, assistant
Comprova	14	873.83	3352.20	47979.00	1	fgts, withdrawal, fund, case, no, day, measure, government, account, also
	13			55.00	1	video, no, vaccination, death, flu, proof, after, report, original, vaccination

International Fact-Checking Network was nominated to the Nobel Peace Prize in 2021 due to its effort in promoting collaborations across countries and organizations for fact-checking[30]. Another challenge for fact-checking organizations is that fact-checkers need, constantly, to choose not only what to debunk, but also when to debunk conspiracy theories and misinformation, avoiding giving undesirable attention to misinformation and conspiracy theories that have not spread. We hope our historical data of fact-checking instances from one specific country can be exploited by studies that attempt to increase the efficacy of fact-checking.

Characterizing and exploring the misinformation that has been debunked in Brazil. In the field of misinformation prevention, fact-checking has observed a tremendous spotlight as an effective tool. In this scenario, the fact-check instances and the agencies responsible for publishing them, become themselves objects of academic interest. The data compiled in our repository may be used to study the fact-checking ecosystem in Brazil. Examples of such analysis might include a study on diverging labels between the agencies for the same misinformation instances or even an analysis of the overlap of topics covered by the fact-checks released by these agencies.

Understanding misinformation in Brazil. It is hard to overstate the magnitude of the damage caused by the unrestricted dissemination of misinformation instances in Brazil. Particularly, this damage was intensified by the COVID-19 pandemic, in which the spread of this kind of content was potentialized by the wide-spread feeling of vulnerability within the Brazilian population and this has amounted to a dramatic scenario in which misinformation is effectively a public health concern. In this context, seeking new perspectives about misinformation in Brazil, such as identifying the key clues and patterns utilized by the fact-checking agencies that allow them to correctly label suspicious instances or even exploring the correlation between the topics covered by published fact-checks and corresponding world events is a primordial step towards the development of new solutions. Furthermore, our repository allows for the cross-referencing between fact-checks and suspicious content, thus allowing for the identification of misinformation instances in the wild. From these instances it is then possible to extract useful spread patterns and also features that might characterize their inception.

[30] `https://bit.ly/3oMtLGa`

Automatic misinformation detection. A frequent challenge in the context of misinformation prevention is the scarcity of data labeled by recognized entities. Our repository might be used to identify instances of misinformation in several suspicious sources. These instances can then be utilized in the training of machine learning algorithms that attempt to identify characteristics and distributions that might be associated with this kind of content. These machine learning models can then be utilized for the automatic detection of misinformation content and the development of tools to assist journalists and fact-checking agencies in debunking misinformation.

ACKNOWLEDGMENT

This work was partially supported by Ministério Público de Minas Gerais (MPMG), project Analytical Capabilities, as well as grants from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

Bessi, A. and Ferrara, E. Social bots distort the 2016 us presidential election online discussion. *First Monday* 21 (11), 2016.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research (JMLR)* vol. 3, pp. 993–1022, 2003.

Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the International Conference on World Wide Web (WWW)*. pp. 675–684, 2011.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. Computational fact checking from knowledge networks. *PLOS ONE* 10 (6): e0128193, 2015.

Córdova Sáenz, C. A., Dias, M., and Becker, K. Assessing the combination of distilbert news representations and diffusion topological features to classify fake news. *Journal of Information and Data Management (JIDM)* vol. 12(1), 2021.

Couto, J. M. M., Pimenta, B., de Araújo, I. M., Assis, S., Reis, J. C. S., da Silva, A. P., Almeida, J., and Benevenuto, F. Central de fatos: Um repositório de checagens de fatos. In *Proceedings of the Dataset Showcase Workshop (DSW/SBBD)*. pp. 128–137, 2021.

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

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., et al. Fake news vs satire: A dataset and analysis. In *Proceedings of the ACM Conference on Web Science (WebSci)*. pp. 17–21, 2018.

Gomes Jr, L. and Frizzon, G. Fake news and brazilian politics–temporal investigation based on semantic annotations and graph analysis. In *Proceedings of the Brazilian Symposium on Databases (SBBD)*. pp. 169–174, 2019.

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

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. The science of fake news. *Science* 359 (6380): 1094–1096, 2018.

Machado, C., Kira, B., Narayanan, V., Kollanyi, B., and Howard, P. A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. In *Companion Proceedings of the World Wide Web conference (WWW)*. pp. 1013–1019, 2019.

Maros, A., Almeida, J. M., and Vasconcelos, M. A study of misinformation in audio messages shared in whatsapp groups. In *Disinformation in Open Online Media*, J. Bright, A. Giachanou, V. Spaiser, F. Spezzano, A. George, and A. Pavliuc (Eds.). Springer International Publishing, Cham, pp. 85–100, 2021.

Martins, A. D. F., Cabral, L., Mourão, P. J. C., Monteiro, J. M., and Machado, J. Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems*. pp. 199–206, 2021.

Moreno, J. and Bressan, G. Factck. br: a new dataset to study fake news. In *Proceedings of the Brazillian Symposium on Multimedia and the Web (WebMedia)*. pp. 525–527, 2019.

Myslinski, L. J. Fact checking method and system, 2012. Google Patents. US Patent 8,185,448.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. Automatic evaluation of topic coherence. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT)*. pp. 100–108, 2010.

Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. K. Reuters Institute Digital News Report 2019. Reuters Institute for the Study of Journalism, 2019.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018.

Reis, J. C. and Benevenuto, F. Supervised learning for misinformation detection in whatsapp. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia)*. pp. 245–252, 2021.

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

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

Resende, G., Melo, P., Reis, J. C. S., Vasconcelos, M., Almeida, J., and Benevenuto, F. Analyzing textual (mis)information shared in whatsapp groups. In *Proceedings of the International ACM Conference on Web Science (WebSci)*. pp. 225–234, 2019.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proceedings of the ACM Web Conference (WWW)*. pp. 818–828, 2019.

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Goga, O., Gummadi, K. P., and Redmiles, E. M. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*. pp. 140–149, 2019.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

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

Tardaguila, C., Benevenuto, F., and Ortellado, P. Fake news is poisoning brazilian politics. whatsapp can stop it. https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html, 2018.

Vlachos, A. and Riedel, S. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*. pp. 18–22, 2014a.

Vlachos, A. and Riedel, S. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*. pp. 18–22, 2014b.

Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. pp. 647–653, 2017.

Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science* 359 (6380): 1146–1151, 2018.

Wang, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

Wu, Y., Agarwal, P. K., Li, C., Yang, J., and Yu, C. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7 (7): 589–600, 2014.