Submission date: 16/10/2021
Resubmission date: 25/02/2022
Resubmission date: 15/05/2022
Camera ready submission date: 02/06/2022

1st round notification: 25/01/2022
2nd round notification: 15/04/2022
3rd round notification: 16/05/2022
Available online: 07/07/2022
Publication date: 07/07/2022

Section: special issue
[Information Systems in Life and Health Sciences]

Fake News Detection in Tweets: Challenges and Adaptations imposed by the COVID-19

Christiane Santana, Daniela Barreiro Claro, Marlo Souza

[1] FORMAS Research Group, Institute of Computing, Federal University of Bahia
Milton Santos Ave., Ondina, 40.170-110, Salvador, Bahia, Brazil

{christiane.santana,dclaro,msouza1}@ufba.br

Abstract. *Misinformation has plagued citizens' lives, especially on social networks. During the COVID-19 pandemic, the proliferation of competing narratives and dissemination of false or inaccurate news about the pandemic has reached such a state that led the World Health Organization to classify it as an infodemic. However, few resources are available to combat misinformation in this new and evolving domain, especially considering how social networks allow the rapid spreading of false narratives. In this case, the lack of resources, such as methods, tools, and reliable information on the virus, hinders our ability to combat this misinformation. In this work, we investigate the application of Text Analysis methods to help health-related scientific communicators produce educational material to combat misinformation. This study was conducted in association with the Scientific Communication sector of FIOCRUZ, a health research institution in Brazil, aiming to monitor COVID-19-related fake news and produce educational material to combat misinformation in a weekly manner due to the ephemeral nature of COVID-19 misinformation in social media. As the main findings of this work, we provide (1) a pipeline for automatically collecting and analyzing news and social media posts regarding COVID-19 in order to provide science communicators with a weekly contextualized view of topics related to COVID-19 in social media; (2) we analyzed the effect of different resources and methods in the analytical tools employed in this work for detecting health-related misinformation in the Portuguese language, and finally, (3) we provided to journalists and science communicators in FIOCRUZ computational tools to automatically monitor COVID-related misinformation in social media, focusing on Twitter, aiming to contribute to definition of the weekly science communication agenda of the institution. Indeed, we indicate the type of resources to combat misinformation in the pandemic, and our approach can handle the detection of misinformation on Twitter social networks within the COVID-19 domain.*

Keywords. *fake news, COVID-19, misinformation, detection*

Cite as: Santana, C., Claro, D. B.
Souza, M. (2022). Fake News Detection in Tweets: Challenges and Adaptations imposed by the COVID-19 (Detecção de Fake News em Tweets: Desafios e Adaptações). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 11:1-11:26. DOI: 10.5753/isys.2022.2286.

1. Introduction

The popularization of social networks and the emergence of Web 2.0 has changed the paradigm of how information is disseminated in contemporary society, enabling the rapid circulation of information. Recently, however, a great deal of attention has been drawn to how such technologies can be exploited to manipulate public opinion and obtain political advantages for certain social groups [Bastos and Mercea 2019][Forelle et al. 2015]. Automated accounts used for mass posting, for example, have become a potential tool to manipulate debates on social media, especially in moments of political relevance. It is estimated, for example, that more than 20% of interactions that took place on Twitter supporting Brazil's general strike in April 2017 were caused by this type of account [Ruediger 2017] and that they also had a significant role in the promotion of misinformation in the United States' 2016 election [Shao et al. 2018, Vosoughi et al. 2018].

People, intentionally or unintentionally, also play an important role in the creation and dissemination of false information in different contexts, such as the electoral one. Cognitive science shows, for example, that "confirmation bias", i.e. people's tendency to believe something that supports their preexisting beliefs, is inherent in human cognition [Plous 1993]. In this context, the study of disinformation, or *fake news*, as it is popularly known, has gained notoriety in recent years, given the massive sharing of false information for political purposes.

Although there are several definitions or classifications for *fake news*, the term is commonly used to label a type of disinformation characterized by the emulation or use of narrative and journalistic styles, with the aim of communicating totally or partially false information to deceive or create misperceptions through its propagation [Tandoc Jr et al. 2018].

In the context of the COVID-19 pandemic, the large volume of information disseminated about this topic on social networks, many of them false or inaccurate, led the World Health Organization to classify the phenomenon as an infodemic [World Health Organization 2020]. Health-related disinformation can have a major impact on people's perception of the risks of COVID-19 and the adoption of health protection measures [Krause et al. 2020, Dryhurst et al. 2020]. As pointed out in studies around anti-vaccine movements [Meleo-Erwin et al. 2017, Schmidt et al. 2018, Yiannakoulias et al. 2019], these discourses arise from disbelief in scientific, media, and political institutions, and that to overcome their spreading, health professionals must understand concerns and discourses surrounding health-related disinformation.

As a recent and still relatively unknown topic, the consumption of information regarding the COVID-19 pandemic has been primarily concentrated within social media [Sharma et al. 2020]. Given several aspects such as the ephemeral nature of sharing information about COVID-19, the revision of sanitary protocols and restrictions, the emergence of multiple waves of the disease in several countries, the constant updates on the local and global state of the pandemic, and the multiplicity of narratives competing for public attention, the need for tools and methodologies that help verifying the facts about COVID-19 and producing reliable scientific communication material is essential.

Thus, our work proposes a methodology for computationally-assisted fact-checking and producing scientific communication materials regarding the COVID-19 pandemic. Our work aims to provide a set of tools for journalists and science communicators to identify COVID-related disinformation and produce informative content to combat harmful narratives. In order to achieve this goal, we employed text analysis and supervised and unsupervised machine learning techniques to identify messages containing fake news in the social network Twitter, based on a database of verified news, as well as the identification of central topics of discussion in a fixed period.

Technically, we evaluate the use of resources, such as automatically-collected COVID-related news articles from trustworthy sources and fact-checked information, and text analysis methods to detect misinformation on tweets datasets provided by experts from FIOCRUZ - a health research institution in Brazil which has been working since 2020 on producing reliable science communication material regarding the COVID-19 pandemic to the Brazilian public[1]. Weekly, FIOCRUZ experts analyze a vast amount of information to provide informative content to combat fake news and, consequently, its dissemination. We evaluate methods of text classification and automatic label propagation to assist them in augmenting their annotated dataset to guide their weekly COVID-related scientific communication agenda.

The rest of this article is organized as follows: **Section 2** presents the scientific communication methodology; **Section 3** describes our detection approach based on text classification; **Section 4** presents the data annotation through label propagation strategy; **Section 5** evaluates our label propagation into text classification; **Section 6** discusses some issues and threats of validity; **Section 7** describes some related work and finally **Section 8** presents our conclusions and future directions.

2. Scientific Communication for combating COVID-19 disinformation

Scientific communication is an essential aspect of scientific practice and involves the communication of scientific ideas and practices both to the specialized and lay public. As such, scientific communication plays an important role in educating the public on scientific practices and interpreting expert opinions on their topics of expertise.

Undoubtedly, Information and Communication Technologies play a crucial role in scientific communication, ever since the invention of the press, and the creation of the Internet allowed increased access and exchange of knowledge in a scale not possible before. Nevertheless, as authors in [Dantas and Deccache-Maia 2020] highlight, technological advances are also accompanied by societal changes in the ways we consume, form relationships, and communicate with each other. In the same way, these societal changes are reflected in how we consume and think critically about scientific ideas and expert opinions on topics of our day-to-day life.

As argued by [van Dijck and Alinejad 2020], social media has had a significant impact on the model of science communication in recent years. While science communication has relied, in the last century, on linear flows of communication between professional actors in academia, media, and decision-making organizations and the public, social media introduces direct channels of communications that defy the previously

[1] https://redecovida.org/

shared assumptions in the whom to trust, in the what to trust and on how this trust is built. It is argued that the shift from the institutional model of communication to a network model, in which every idea, whether from experts or non-experts, are treated as resources in a “marketplace of ideas”, has led to the erosion of trust in institutions, such as science, education, government and news agencies [Dahlgren 2018].

Observing the exchange of information regarding the pandemic in the Netherlands during the first four months of the outbreak, authors in [van Dijck and Alinejad 2020] identify that social media can be used both to build trust and distrust. At the same time that the proliferation of discourses and points of view may reduce trust in expert voices and institutions, it allows citizens to contrast their realities with official reports and provide greater transparency and accountability to decision-makers, building trust in official accounts. The authors conclude that the mechanisms of communicating reliable information are central aspects of the epidemic’s evolution and that social media can act both as conduits for misinformation and scientific information to the public.

According to Messeder Neto [Messeder Neto 2019, apud Dantas and Deccache-Maia 2020], in the current context, for the scientific communicator, it is necessary not only to present accurate scientific knowledge but actively fight pseudo-scientific ideas and scientific disinformation. It is not enough to show what is correct but also to point out incorrect information in our networks.

In this sense, as pointed out by [Jacobsen and Vraga 2020], there are specific actions that health professionals can take to reduce misinformation about COVID-19 and disseminate accurate, evidence-based information to the public. Particularly, the authors point out the importance of health communication specialists in countering prominent false narratives usually propagated in social media and communicating reliable information in a contextualized manner. This is particularly important since, as reported by [Vraga and Bode 2017], authoritative sources can have a significant role in combating health-related misinformation in social media.

This work is focused on applying text analysis methods for the Portuguese language to help health professionals and scientific communicators to fight (mis)disinformation on COVID-19 in social media. We employ different machine learning techniques for detecting COVID-related fake news, monitoring discourses and narratives in the public sphere that need attention from health professionals, and building scientific communication materials to combat the spreading of false or misleading information about the new coronavirus and the pandemic.

As Figure 1 shows, the first step in our scientific communication pipeline refers to monitoring pandemic-related content on social media. In this work, we achieved this by a continuous collection of messages related to the pandemic published on social media, which were later analyzed by our tool. The next step takes the fake news detection and the relevant topic identification on misinformation during a week. These topics were ranked as the most exciting topics of the week. The dissemination of communication material is conducted by scientific communicators and health professionals and publicized for over a week.

isysTemplateLatex/SBC - iSys Template/images/Pipeline.png

Figure 1. COVID-related disinformation combat and scientific communication pipeline

The following sections describe our methodology to detect fake news to help the production of communication material. This work results from a partnership with FIOCRUZ in combating misinformation during the pandemic.

3. COVID-related Fake News detection through Text Classification

Fake news detection is a new area of study that aims to assess the factuality of a given piece of text, i.e., determining whether (to what degree) the information contained in the text is true. It has gained large interest in the last five years due to the increasing circulation of misinformation in social media [Pérez-Rosas et al. 2018, Marín and Arroyo 2019, Oshikawa et al. 2020]. Despite its simplicity and conceptual limitations, approaches for fake news detection based on classification considering simple linguistic features, such as lexical and stylistic characteristics of deceptive language use, have achieved considerable accuracy in the available datasets [Oshikawa et al. 2020, Al-Rakhami and Al-Amri 2020]. These models have achieved good performances for the Portuguese language on corpora such as the Fake.Br [Monteiro et al. 2018, Ruiz and Okano 2019, Silva et al. 2020] and FakeTweet.Br [Cordeiro and Pinheiro 2019].

Health misinformation has a different nature from politically-motivated disinformation as asserting the truthfulness of health-related information concerns verifying scientific claims, which is not an easy task - as recognized by authors such as [Wadden et al. 2020, Paka et al. 2021]. However, it is unclear how useful previous models for fake news detection based on linguistic and stylistic information can be to detect COVID-related fake news and help monitor the discourse surrounding the pandemic on social media. Some recent work [Paka et al. 2021, Al-Rakhami and Al-Amri 2020, Almatarneh et al. 2021, Priya and Kumar 2021] employed supervised/semi-supervised approaches to detect COVID-19 misinformation. However, the effort to label data and the linguistic peculiarities of social media texts, particularly on Twitter, have encouraged some to link tweets with external knowledge to improve accuracy. Considering the effort to label data, almost all of these works employ static corpora, not dealing with the time-sensitive nature of COVID-related fake news, which would require continuous data annotation and model updating each time the narratives evolve concerning a particular topic, such as treatments, means of propagation, and the vaccine. In this work, we made an effort to provide a label propagation approach for auto-supervising learning. Indeed, considering the scarcity of computational linguistic resources for the Portuguese language these limitations in the literature for fake news in the English language are further aggravated when dealing with fake news in Portuguese.

In the following, we evaluate the use of different resources, such as general domain and domain-specific fake news corpora, as well as different text representation models, namely distributional models, to detect COVID-related misinformation in social media.

Our classification modeling of the problem of fake news detection based on linguistic features is composed of three steps. First, each social media message or piece of news to be verified is preprocessed by tokenization and computing semantic vector representations of the text, using one of the investigated representation models. These vector representations are then fed, in a feature-based style, to the classification model that classifies the message as truthful (FACT) or false (FAKE), or undecided (DISCARD). The overall pipeline of our classification approach is depicted in Figure 2.

isysTemplateLatex/SBC - iSys Template/images/Figura2-pipeline.png

Figure 2. Classification approach to Fake News Detection

In this work, we focused on messages from the Twitter social network due to its importance in the spread of information (and disinformation) regarding the COVID-19 pandemic and the technical ease of obtaining data through the platform's API.

Our first hypothesis was that **(1)** it would be possible to re-use the datasets for fake news detection to detect misinformation tweets related to COVID-19, exploiting linguistic cues for detecting misinformation, even considering the difference among textual styles between the journalistic datasets and health-related tweets. We test this hypothesis due to the scarcity of datasets for COVID-related fake news detection in social media and the ephemeral nature of pandemic-related misinformation, which quickly renders such datasets irrelevant from a practical standpoint.

Our methodology analyzes the available resources, domains, and writing styles to detect disinformation about COVID-19 circulating on Twitter in the Portuguese language. It considers different distributional semantic representation models that can be used for fake detection in tweets for the Portuguese language.

3.1. Distributional semantic representation models

Distributional semantic models (DSM) are based on the assumption that the meaning of a word can be inferred from its usage, i.e., its distribution in text. Therefore, these models dynamically build semantic representations, in the form of high dimensional vector spaces, through a statistical analysis of the contexts in which words occur. Distributional semantic representation models are a technique for solving the lexical acquisition bottleneck by unsupervised learning. Their distributed representation provides a cognitively plausible, robust, and flexible architecture for the organization and processing of semantic information [Evert 2010].

Distributional semantics is based on the Distributional Hypothesis, which states that similar meaning is similar to linguistic distribution: Words that are semantically related, such as *post-doc* and *student*, are used in similar contexts [Boleda 2020]. It represents word meaning by taking large amounts of text as input and, through an abstraction mechanism (symbolized by the arrow in Figure 2), producing a distributional model akin to a lexicon, with semantic representations in the form of vectors. The vectors for *post-doc* and *student* are closer in the space than those of *postdoc* and *wealth*, because their vector values are more similar.

Three existing distributional semantics models were evaluated: *Bag-of-words* with TF-IDF, *Word2Vec* [Mikolov et al. 2013] and *DistilBERT* [Sanh et al. 2019]. Each text collected from tweets was tokenized and vectorized using each model. We employed a pre-trained *Word2Vec*[Mikolov et al. 2013] with skip-gram for the Portuguese language with 300 dimensions [Hartmann et al. 2017] and a pre-trained *DistilBERT*[Sanh et al. 2019] for the Portuguese language with with 300 dimensions.

3.1.1. Bag-of-Words

The first distributional representation was the *Bag of Words*. Each news or tweet is treated as a set of words, regardless of the order and context in which they appear, thus generating a term-document matrix.

Each word frequency in *Bag of Word* considers the number of times it is found in the text. Using the TF-IDF (*Term Frequency Inverse Document Frequency*), a weight is assigned to each term, pondering its frequency. The more frequent in distinct texts a term is, the lower its weight, thus making it possible to consider the relevance of the term in a classification process.

One of the advantages of this approach is that it is language-independent. However, it does not consider the semantic relations between the terms in the vocabulary, and the generated matrix is very sparse.

3.1.2. Word2vec

Word2Vec [Mikolov et al. 2013] is a distributional model based on shallow neural networks where each word is represented as a fixed-size feature vector, obtained from a pre-trained model, establishing semantic relationships between the terms.

Word2Vec allows the use of two models: CBOW (*Continuous Bag of Words*), which is based on a context and suggests the central word. Meanwhile, Skip-Gram, the model employed in this work, suggests a particular context from a central word. In this case, the context is defined by the word that precedes and follows the central word.

3.1.3. DistillBERT

DistilBert [Sanh et al. 2019] is a distributional representation of deep, contextualized, lightweight language model obtained by distilling a BERT language model [Devlin et al. 2018]. As this method is a contextualized representation model, the representation of words depends on their syntactic context. The model can represent contextual variations in the semantics of lexical units.

After computing vector representations for the texts, a supervised machine learning method generates a classification model to predict fake news. In our experiments, we employed the SVM classification algorithm [Lorena and Carvalho 2003] with a linear kernel, which presents a satisfactory performance in textual classification tasks [Joachims 2002].

In the following sections, we describe each strategy to detect fake news on tweets for COVID-19.

3.2. Experimental Setup

The first investigation on applying different resources and methods for COVID-related fake news detection was evaluating the effects of knowledge domain and textual styles in the corpora in predictive performance of the fake news detection methods.

For that, we employed three different corpora to train and test our models, analyzing the impact of text-domain and style in detecting fake news. In this first attempt, we employed three different datasets: *FAKE.BR*, *G1FatoFake* and *TweetsCOVID*.

The *FAKE.BR* corpus, created by [Monteiro et al. 2018], is composed of 7196 sentences labeled as truthful or false, obtained from newspaper articles on various topics, and standardized to minimize biased results. This database is balanced, containing 50% of trustworthy news and 50% of false news, treated in pairs. That is, for each true sentence on a given subject, there is a false correspondent on the same topic. This dataset can be characterized as general domain, i.e. involving news about diverse topics such as politics, economics, art, etc., with a journalistic writing style.

The *G1FatoFake* corpus was compiled by us to generate a database of journalistic news related to COVID-19. Thus, approximately 200 manually fact-checked news items were retrieved from the portals G1 – Fato or Fake and G1 – Bem Estar, all referring to the COVID-19 domain, being 50% truthful and 50% false. The collection of articles was carried out via *web scrapping*, selecting only the news that refer to the themes “COVID” and “coronavirus” and classified as truthful (FACT) or false (FAKE). In this experiment, we employed only the titles of the news, as they were closer in size of the messages collected on Twitter, aiming to reduce bias the results. This dataset is characterized by the COVID-19domain and journalistic writing style.

The *TweetsCOVID* corpus was generated from COVID-19 messages collected on Twitter by FIOCRUZ, between January 26 and 28 and February 2 and 4, 2021. It was collected approximately 72000 tweets. However, of these, only 19 messages were labeled by FIOCRUZ specialists as FACT or FAKE due to the difficulties encountered in manually checking the veracity of the information.

Concerning the *TweetsCOVID* corpus, the 19 annotated records were considered, 16 as false (FAKE) and 3 as truthful (FACT). This dataset is characterized by the COVID-19 domain and Twitter writing style. Table 1 summarizes our corpora by style written and domain.

	Corpora		
	FAKE.BR	G1FatoFake	TweetsCOVID
Number of lines	7196	197	19
Style	News	News	Tweets
Domain	General	COVID	COVID

Table 1. Our corpora

Considering each corpus, we proposed three experiments varying, for each of them, the corpora for training and testing the model to evaluate our hypothesis, as described in Table 2.

Experiment	Train	Test	Written Style	Domain
1	FAKE.BR	G1FatoFake	News x News	General x COVID
2	FAKE.BR	TweetsCOVID	News x Tweets	General x COVID
3	G1FatoFake	TweetsCOVID	News x Tweets	COVID x COVID

Table 2. Experiments per written styles and domains

3.3. Results

We provide our results considering each experiment from Table 2.

3.3.1. 1^{st} Experiment: Similar Styles and Different Domains

In the first experiment, the corpus *FAKE.BR* was the training set within a general domain and journalistic style, and the test set was *G1FatoFake*, with the same style but within the COVID-19 domain.

Vector Representation	Precision	Recall	F1
BOW	0.51	0.71	0.59
Word2Vec	0.59	0.70	0.64
DistilBERT	0.63	0.58	0.60

Table 3. Performance of the Supervised Model on the 1^{st} Experiment

Comparing the results, it is possible to observe that the method *Word2Vec* obtained the best performance, as shown by the data presented in Table 3. Notice that in our results, the models trained over the FAKE.BR corpus have presented bias towards classifying COVID-related texts as fake news, possibly due to the lexical and style differences in health-related news to other domains such as art, economics and politics.

3.3.2. 2^{nd} Experiment: Different Styles and Domains

In the second experiment, the goal was to verify the impact on employing different datasets with different textual styles and domains on training and testing. Thus, the *FAKE.BR* corpus was kept as a training set, thus remaining the general domain and journalistic style, and the adopted test set was the *TweetsCOVID*, with COVID-19 domain and tweets style.

Vector Representation	Precision	Recall	F1
BOW	0.20	0.68	0.31
Word2Vec	0.0	0.0	0.0
DistilBERT	0.0	0.0	0.0

Table 4. Results from the 2^{nd} Experiment

The experiments with Word2Vec and DistilBERT representations, presented in Table 4, did not yield results due to the small number of annotated examples in the test dataset TweetsCOVID, composed of 19 instances manually annotated by FIOCRUZ experts, because of the hard manual labeling task. This fact encourages the investment in automatic or semiautomatic methods for the construction of robust datasets to train a classification model to be employed in the pipeline of detecting misinformation in social media messages, by the prediction of non-labeled messages.

3.3.3. 3^{rd} Experiment: Different Styles and Similar Domains

In the third experiment, the training corpus became *G1FatoFake*, so the domain became COVID-19 specific, and the journalistic style and test set used was *TweetsCOVID*, with the same domain - COVID-19 and different style - tweets.

The 3rd experiment yields similar results to that of the 2nd, due to the size of the test dataset.

Vector Representation	Precision	Recall	F1
BOW	0.20	0.68	0.31
Word2Vec	0.0	0.0	0.0
DistilBERT	0.0	0.0	0.0

Table 5. Performance of the Supervised Model on the 3^{rd} Experiment

In all experiments, better results were achieved by the *Bag of Word* distributional representation. The other approaches do not achieve a result due to the number of records to perform. In this context, our second attempt to improve the detection of fake news was to provide a mechanism to annotate tweets with large number of data. In this way, we proposed a label propagation approach as described in the following section.

4. Data Annotation through Label Propagation

One of the main concerns regarding COVID-related fake news is the fast pace at which new narratives arise and disappear in social media. Also, new scientific discoveries and relevant political events occur daily and must be considered when evaluating the trustworthiness of the information. As such, any fake news detection model needs to be constantly updated with new information.

However, it is unfeasible to manually collect and label examples of truthful and deceptive messages on time to update these models. As such, we propose a semi-supervised method of fake news detection, employing information from trustworthy sources, such as journalistic and fact-checking agencies, to label social media messages through label propagation, which can then be used to re-train the classifier incorporating new information.

Thus, our second hypothesis was that **(2) it would be possible to improve COVID-related fake news detection if we augment the tweets datasets**.

The label propagation method proposed in this work relates each social media message with a set of trustworthy information, such as news articles from trusted journalistic sources and information previously verified by fact-checking agencies that either support or contradict it. In order to perform this analysis, we investigate the application of semantic similarity measures to identify which of the trusted sources can be applied to verify the message.

To carry out this label propagation method, it was necessary to create a knowledge base with COVID-related news and information labeled as truthful or false, or simply FACT or FAKE. Thus, for this work, a database of journalistic news related to COVID-19 was collected from the fact-checking agency G1 – Fato ou Fake [2] and health-related portal G1 - Bem Estar [3] via web scrapping from November 2020 to April 2021 . This dataset, called *CHECKED_BASE*, is composed of four fields: *Label, Title, Summary and Link*, as described in Table 6 and by 200 annotated news items, 99 classified as *FAKE* and 101 news classified as *FACT* as presented in Figure 3. To create this knowledge base, only the

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

[3] https://g1.globo.com/bemestar/

isysTemplateLatex/SBC - iSys Template/images/ExemploBaseChecagem.

Figure 3. Example of a CHECKED_BASE instance

news headlines were considered to maintain size compatibility with the tweets, reducing the risk of bias in the results. With the compiled base, an experiment was carried out for the propagation of labels using semantic similarity based on the previously discussed distributed representation models to assert when given trusted information can be used to verify a tweet.

Column	Description
Label	the classification of the news article as either truthful (FACT) or false (FAKE)
Title	The title of the news article
Summary	A summary of the news article
Link	An access hyperlink to the news article

Table 6. Schema for the CHECKED_BASE

4.1. Semantic Similarity

Semantic similarity measures try to estimate how two words or text fragments are close to each other in regards to their meanings and have a great range of applications from information retrieval to word sense disambiguation [Li et al. 2003]. One application of textual semantic similarity measures of particular interest to us is in the task of label propagation [Pawar et al. 2017, Wang et al. 2011, Dumitrache et al. 2018]. Label Propagation (LP) is the task of assigning labels to unlabeled points in a dataset by propagating them from other previously labeled points [Zhu and Ghahramani 2002]. In this task, semantic similarity measures can be useful to select appropriate labels for unlabeled points, by selecting labeled data which is semantically close to them.

Our approach aims to propagate reliable information, such as articles curated from reliable news outlets and fact checkers, obtained directly from G1 News (Fato ou Fake and Bem Estar), which we call the CHECKED BASE dataset, into the non-annotated tweets through an automatic alignment method. With this, we aim to semi-automatically construct a large corpus with an existing knowledge of known facts and annotated misinformation.

4.2. Experimental Setup

Our semantic similarity measure is based on textual vector representations obtained from our distributional representation models. Particularly, we employed the aggregation of DistilBERT word representations for each word in the text, since DistilBERT can capture context information in its representation. To measure the similarity between two texts, we employed the cosine similarity between their vector representations. To propagate the labels from the *CHECKED BASE* to the *tweets corpus*, we compute the semantic similarity between each collected tweet and the title of each news article in our *CHECKED BASE*. As before, we employ only the title of the news in our experiments to reduce noise due to size differences between the texts. Each tweet received the label from the most similar news article in the *CHECKED BASE*, if their similarity was greater than a certain threshold. We adopted the threshold of 60% similarity - established empirically with the FIOCRUZ experts in preliminary experiments with validation data not included in these experiments.

We used the tweets base, TweetsPerWeek15, generated by FIOCRUZ with tweets collected from April 12 to 15, 2021, corresponding to the fifteenth week of tweet monitoring carried out by the Institution to carry out the propagation of labels. The dataset provided by FIOCRUZ is composed of 35169 records. After pre-processing these data, duplicated and short messages (with up to 5 words) were excluded, and the number of records was reduced to 7389. This pre-processed database was submitted to the label propagation process. A new corpus of tweets was generated, named *ANNOTATED_TWEETS* with three possible labels: *FACT* or *FAKE* to those tweets with a semantic similarity above the threshold to some verified information in the *CHECKED_BASE*, and *NOT ANNOTATED* for those tweets for which no verified information could be related with similarity above the threshold, and thus, no label was propagated.

4.3. Results

Analyzing the results of the label propagation process, it is possible to observe that the majority of the tweets, almost 58%, were annotated as FACT, and a few more than 25% were annotated as FAKE. Only 17% of tweets were not labeled by our method, thus assigned the NOT ANNOTATED label, as can be observed in Table 7.

Dataset	FAKE	FACT	NOT ANNOTATED
TweetsPerWeek 15 ANNOTATED_TWEETS	1876 (25,39%)	4279 (57,91%)	1234 (16,70%)

Table 7. Results from label propagation by similarity

This new annotated dataset *ANNOTATED_TWEETS* obtained from our label propagation approach was manually validated by three experts from FIOCRUZ. They provide the evaluation based on a set of features: (i) tweet; (ii) label propagated; (iii) news title aligned with the label propagated; (iv) website link from the news and (v) degree of similarity.

Experts from FIOCRUZ employed new labels to validate the annotation of the dataset (*ANNOTATED_TWEETS*). They employed FACT or FAKE to those true or false messages which were possible to validate, respectively. They employed the UNDETERMINED for those analyzed but not classified due to lack of evidence. The NA (Not Applicable) was used to refer to tweets considered by them that were not possible to be classified, such as expressions of opinions, questions, and other non-declarative messages. The experts evaluated 6155 labeled tweets, disregarding those noted as NOT ANNOTATED in dataset *ANNOTATED_TWEETS*. These tweets not assessed by the experts were labeled as DISCARD.

Within the new results by FIOCRUZ experts, it is possible to observe that just over 50% of the tweets were labeled, with approximately 4.6% classified as FAKE and 45.6% labeled as FACT, as described in Table 8. The other set of tweets was not analyzed (16.7% - DISCARD); the other set of tweets was not classified due to lack of evidence (8.4%- NOT LABELED) or the divergence of categorization (24.64%-NA) due to the nature of the message, the divergence of opinions or ideas among the annotators. We called this new dataset as (*FIOCRUZ_TWEETS*)

Dataset	FAKE	FACT	DISCARD	NA	NOT LABELED
TweetsPerWeek 15 *FIOCRUZ_TWEETS*	342 (4,63%)	3370 (45,61%)	1234 (16,70%)	1821 (24,64%)	622 (8,42%)

Table 8. Results from FIOCRUZ experts at *FIOCRUZ_TWEETS*

All tweets that could not be classified as FACT or FAKE were considered NOT ANNOTATED. Thus, the Table 9 shows the distribution of labels in the dataset (*FIOCRUZ_TWEETS*).

Dataset	FAKE	FACT	NOT ANNOTATED
TweetsPerWeek 15 *FIOCRUZ_TWEETS*	342 (4,63%)	3370 (45,61%)	3677 (49,76%)

Table 9. Results from FIOCRUZ experts at *FIOCRUZ_TWEETS*

Table 10 compares the results from *ANNOTATED_TWEETS* and *FIOCRUZ_TWEETS*. It is noticeable a reduction either from FAKE and FACT labeled tweets. The FACT tweets reduced 12.3% and the FAKE tweets reduced more than 20%.

Dataset	FAKE	FACT	NOT ANNOTATED
TweetsPerWeek 15 *ANNOTATED_TWEETS*	1876 (25,39%)	4279 (57,91%)	1234 (16,70%)
TweetsPerWeek 15 *FIOCRUZ_TWEETS*	342 (4,63%)	3370 (45,61%)	3677 (49,76%)

Table 10. Manual Annotation x Label Propagation Annotation

Table 11 indicates that the assertiveness rate of manual and label propagation annotations was just over 40%. Although we consider a low measure for the experiment, the degree of similarity adopted (threshold of 60%) was considered a unique criterion.

It is possible to observe that of 6155 automatically labeled tweets, 2443 (about 40%) were no longer categorized by the experts due to lack of evidence to prove the assigned label or, because they are incomplete, out of context, or because they are not news, but opinions not subject for factuality judgments. Among 4279 tweets classified as FACT by the label propagation method, 2344 (approximately 55%) were validated by the human analysis; 208 (4.86%) were changed to FAKE and 1727 had the label revoked, due to the nature of the message or due to disagreement among annotators. Among 1876 classified as FAKE, only 134 (about 7%) were kept after manual validation; 1026 (about 55%) had their labels changed to FACT, and 716 were no longer classified.

Manual Annotation	Label Propagation Annotation		
	FAKE 1876	FACT 4279	Total of Annotated 6155
Label confirmed	**134 (7,14%)**	**2344 (54,78%)**	**2478 (40,26%)**
Label modified	1026 (54,69%)	208 (4,86%)	1234 (20,05%)
Label revoked	716 (38,17%)	1727 (40,36%)	2443 (39,69%)

Table 11. Comparative: Manual Annotation x Label Propagation Annotation

With this new dataset *FIOCRUZ_TWEETS* we provide a new set of experiments to improve fake news detection as described in the following sections.

5. Fake news detection through Social Networks with Label Propagation

The labeling propagation process enable us to envision a new set of experiments to detect fake tweets by a classification approach. In the following subsections we describe our methodology to validate our second hypothesis on the improvement of detection of fake tweets.

5.1. Experimental Setup

The *FIOCRUZ_TWEETS* corpus was generated from information about COVID-19 collected on Twitter as described in the previous section. The 3712 annotated records were considered, 342 as FAKE and 3370 as FACT, with this database being unbalanced. This dataset is characterized by the COVID-19domain and Twitter writing style. Table 12 summarizes our corpora by style written and domain.

	Corpora		
	FAKE.BR	G1FatoFake	*FIOCRUZ_TWEETS*
Number of lines	7196	197	3712
Style	News	News	Tweets
Domain	General	COVID	COVID

Table 12. Our corpora

Considering each corpus, we proposed four new experiments as described in Table 13. Experiments 4 and 5 provides the same experiments 1 and 2 respectively, only modifying the dataset of tweets, now employed the *FIOCRUZ_TWEETS* corpus.

Exp	Train	Test	Written Style	Domain
4	FAKE.BR	*FIOCRUZ_TWEETS*	News x Tweets	General x COVID
5	G1FatoFake	*FIOCRUZ_TWEETS*	News x Tweets	COVID x COVID
6	*FIOCRUZ_TWEETS*	*FIOCRUZ_TWEETS*	Tweets x Tweets	COVID x COVID
7	*ANNOTATED_TWEETS*	*FIOCRUZ_TWEETS*	Tweets x Tweets	COVID x COVID

Table 13. Experiments per written styles and domains

Experiment 6 provides a cross-validation approach to better analyze the *FIOCRUZ_TWEETS* dataset concerning its precision and recall. We performed cross-validation with 5-folds and stratified because this dataset is unbalanced. Experiment 7 analyzes the prediction capability of the *FIOCRUZ_TWEETS* over an annotated dataset *ANNOTATED_TWEETS* to evaluate their prediction. Results are described in the following subsections.

5.2. Results

5.2.1. 4^{th} and 5^{th} Experiments: Styles and Domains

The 4^{th} experiment employed the *FAKE.BR* corpus as a training set remaining the general domain and journalistic style; the test set was *FIOCRUZ_TWEETS* with COVID-19 domain and tweets style. We are using the macro average between FACT and FAKE.

The 5^{th} experiment performed the *G1FatoFake* as training with the COVID-19 domain and the test set was *FIOCRUZ_TWEETS* with either COVID-19 domain. They explored the same domain (COVID-19) and different style (news x tweets).

Regarding the adaptation of style and domain, it is noted that in experiment 1, which explored corpora with the same styles (newspapers), the results achieved were better than in experiments 4 and 5, which worked with different styles, training the models

Vetor	Precision	Recall	F1
BOW	0.54	0.61	0.43
Word2Vec	0.56	0.68	0.51
DistilBERT	0.55	0.64	0.49

Table 14. Results of the 4th experiment with model trained over the FAKE.BR corpus and tested on the FIOCRUZ_TWEETS corpus

Vector Representation	Precision	Recall	F1
BOW	0.52	0.52	0.52
Word2Vec	0.45	0.50	0.48
DistilBERT	0.45	0.50	0.48

Table 15. Results of the 5th experiment with model trained over the G1FatoFake corpus and tested on the FIOCRUZ_TWEETS corpus

with journalistic news and testing with Twitter posts. Two factors can explain this: the first factor is related to the test set used in this work, which contains less than 10% of the records classified as FAKE, thus presenting an imbalance between the classes. Furthermore, the approaches adopted are dependent on lexical and syntactic information captured by the (contextual) representations employed. The second factor is that Twitter messages differ significantly in size and style from news reporting in the training corpora. Thus, classifiers tended to classify Twitter messages as truthful. These experiments indicate that solutions for detecting fake news in short text messages, such as tweets, must be designed, considering their linguistic specificities.

It is noticeable from the results that the domain of the corpus had small impact on the results. This may indicate that the textual style had more significant influence on the results.

5.2.2. 6^{th} Experiment: Same dataset with cross-validation stratified

In this sixth experiment, we trained and tested with the same corpus *FIOCRUZ_TWEETS*. However, we divided it into 5-folds in a cross-validation approach, and we performed the stratification due to an unbalanced dataset.

Vector Representation	Precision	Recall	F1
BOW	0.90	0.54	0.56
Word2Vec	0.55	0.50	0.48
DistilBERT	0.45	0.50	0.48

Table 16. Results with *FIOCRUZ_TWEETS* with cross-validation stratified

From the results presented in Table 16, it is possible to observe that using a training base in the same textual style of the test base, the model achieved better classification precision, especially for the BOW representation. However, the model has obtained no increase in the recall measure, mainly because the corpus is highly unbalanced, with only around 10% of examples labeled as fake.

5.2.3. 7^{th} Experiment: Evaluating the prediction capability

In this last experiment (Table 17), we still perform a 5-fold cross-validation with the *ANNOTATED_TWEETS* and performed the test with the *FIOCRUZ_TWEETS* to evaluate the prediction capability of the model generated by the classification approach.

Vector Representation	Precision	Recall	F1
BOW	0.52	0.54	0.51
Word2Vec	0.45	0.50	0.48
DistilBERT	0.45	0.50	0.48

Table 17. Evaluation measures for fake news detection on the model trained on the base annotated through label propagation and testes on the manually annotated tweets

6. Discussion and Threats of Validity

Information related to COVID-19 is current and ephemeral, in the sense that the most relevant news in circulation has a relatively short life. Thus, it is important that any model trained for the problem of detecting fake news relating to this topic is robust in relation to the time the publication. Thus, to avoid spurious correlations in the experiments performed, the COVID-19domain corpora were compiled within the same period.

While the classification method has proven mildly successful to identify instances of fake news based on linguistic and stylistic cues, our experiments with label propagation indicate that it is possible to integrate reliable and up-to-date factual information on the models, since the method was able to identify truthful information with greater assertiveness.

In addition, it was also evidenced that the pre-processing step of the tweets base needs to be refined, since the characteristics of this type of message make the semi-automatic labeling process very difficult due to informality and lack of standard in the language use. Further, social media messages present the difficulty that they can contain declarative, thus subject to factuality judgements, and non-declarative contents, such as opinions, questions, etc., not being subject to value judgment. Thus, during pre-processing, it is important that this type of message, as well as those that are incomplete or out of context, are identified and excluded, maintaining only those susceptible to categorization in the database to be submitted to the labeling process.

It is important to notice that fact checking is a complex task that is subject to critical epistemological considerations [Uscinski and Butler 2013]. Nevertheless computational methods can be used to help journalists and communicators in performing such a complex task [Ciampaglia et al. 2015]. It is, thus, necessary to understand our work as an effort to contribute with tools to the problem of COVID-related science communication, based on a human-in-the-loop [Zanzotto 2019] architecture, in which our results serve as support for health professionals and science communicators in delineating better approaches to inform the public about the pandemic. In this context, our results are relevant to help them in their manual weekly analysis, as ranking relevant topics to monitor can be supported through the label propagation method, which is useful to identify popular topics with high factual content in social media (as seen in Table 11). Coupled with the classification methods, this enables us to better separate topics with competing narratives that the experts should further examine, automating a previous entirely manual task.

7. Related Work

The task of disinformation detection has been broadly disseminated, but it has mainly exploded in the last two years. Considering the pandemic, misinformation concerning COVID-19 and related areas has imposed urgent actions to stop fake news dissemination.

Until the beginning of the pandemic, most works were focused on the production of corpora and methods to detect fake news based on textual features to indicate a potential misleading content, such as explained in [Silva et al. 2020].

Recently, two literature reviews of the area were conducted by [Giordano et al. 2020] and [Medeiros and Braga 2020]. The first studied different machine learning methods in the literature for fake news detection [Giordano et al. 2020]. The authors fails to provide a consistent evaluation on these methods, and their findings are not conclusive as a critical analysis of the area.

Authors in [Medeiros and Braga 2020] provide a systematic review to determine an overview of the area summarizing the literature on fake news detection until 2019. Due to the recent interest and sudden development of the area of automatic fake news detection on social media, this work does not address the most recent advances in the area, however it is important to notice some conclusions drawn by the authors. First, the authors notice that there has been a great focus of research in Twitter and Weibo, given their popularity and how they are used to spread news by their users. Considering the machine learning techniques used in the literature, they notice that most works employ some neural networks to detect fake news. Regarding the datasets employed in the area, most work propose their own datasets, which indicates a methodological immaturity of the area.

Considering the Portuguese language, authors from [Silva et al. 2020] employed a set of supervised methods within three different vector representations (BoW, Word2vec, and FastText). They described the pipeline to build a corpus of fake news (Fake.BR corpus) following some criteria, such as the length of a sentence, balanced disinformation and information as positive and negative instances, and public available. Such corpus was one of the corpora employed in this work to evaluate our approach. They evaluate some linguistic aspects and feature-based vector representation through three techniques (BoW, Word2Vec, and Fasttext). As far as the pipeline proposed by [Silva et al. 2020] was similar to ours, we concentrated our efforts on the COVID-19 domain and in Twitter texts, which increased the challenge of detecting fake news. More recent work focused on the Portuguese language ([Cruz et al. 2021]) investigated the use of supervised approaches with a labeled corpus to detect COVID-19 fake news. The authors manually annotated their data and extracted a set of features provided to the classifier. The authors do not mention the dynamicity of fake news in social media, especially for politically-motivated fake news, over their static labeled corpus. Although authors [Cabral et al. 2021] created a Portuguese corpus for fake news based on *Whatsapp*, they evaluate their method by considering the extraction of features. Our approach increased the challenge of detecting fake news on Twitter because of the text within the noisy and very limited length.

Working with Fake news on the COVID-19 domain, we observed some related works. Authors in [Anoop et al. 2020] provided a method to enrich emotion information within documents by leveraging emotion lexicons. They provide a relationship between emotions and health-related fake news and a preliminary qualitative on the COVID-19 dataset. The authors indicate that emotion-oriented techniques are a potential direction for tackling COVID-related fake news. While, their results on COVID-19 fake news have great potential, they state that no large fake news dataset at that moment was available to perform some experiments.

Authors in [Zhou et al. 2020] provide a multimodal repository of COVID-related news and their associated levels of credibility. Further, the authors validate the usefulness of their data through preliminary experiments, training baseline classifiers on their data to predict the credibility of COVID-related news. While this resource is of great value for the area, however, the authors also do not deal with the ephemerality of COVID-related information.

The work of [Al-Rakhami and Al-Amri 2020] is the closest work to ours since they aim to detect COVID-related disinformation in tweets. They evaluated machine-learning algorithms trained on a manually annotated dataset with on tweet-based and network-based features. Similarly, authors in [Almatarneh et al. 2021] and [Priya and Kumar 2021] analyzed the use of supervised classifiers for detecting COVID-19 fake news on previously annotated data. While network-based features are important characteristics of social media data for detecting fake news, as it is well-known that social bots are often employed in fake news dissemination campaigns [Shao et al. 2017, Wang et al. 2018, Jones 2019], these methods are limited in scope. First for being opaque to science communicators and health professionals, and second for COVID-related misinformation may be spread organically in the network, not only with malicious intent. As such, content-based techniques are still relevant.

A new recent method [Paka et al. 2021] achieves better accuracy in detecting fake news with a semi-supervised approach. The authors build a manually labeled dataset and train a neural architecture that employs external knowledge and network features to detect fake news. By employing external knowledge, namely the first result in a google search based on the tweet's content, the authors expect their model to be able to keep up-to-date with the evolution of the domain but do not conduct any empirical validation of this hypothesis. Our approach deals with the dynamicity of the fake news domain by employing label propagation to generate data to train our models continuously.

The main difference in our work lies in our focus on employing trustworthy information to avoid continuous manual annotation, which allows our models to incorporate up-to-date information on the pandemic and competing narratives in the public sphere.

As such, we propose a content-based method for training and updating health-related fake news detection models through label propagation and text classification.

8. Conclusion and Future Work

This work proposes a methodology for computationally assisted fact-checking and production of scientific communication materials regarding the new coronavirus and the COVID-19 pandemic, employing Text Analysis and semi-supervised machine learning methods to identify fake news and deceitful discourses about COVID-19 in social media. Our work aims to provide tools for journalists and scientific communicators to identify COVID-related disinformation and produce informative content to combat harmful narratives.

Results presented here point to possible future improvements in our methods, especially in the label propagation method, which we deem crucial to applying the proposed methodology in practice. In the future, we aim to investigate the application of different semantic similarity measures to increase the assertiveness of our method. Notably, we believe that methods based on Textual Entailment Recognition [Dagan et al. 2005, Fonseca et al. 2016] and Neural Information Retrieval [Mitra et al. 2018] are promising avenues of research to improve such results.

Acknowledgement

We would like to thank FIOCRUZ with the Project "Divulgação Científica no Enfrentamento das Fake News durante a Pandemia do COVID-19".

References

[Al-Rakhami and Al-Amri 2020] Al-Rakhami, M. S. and Al-Amri, A. M. (2020). Lies kill, facts save: Detecting covid-19 misinformation in twitter. *IEEE Access*, 8:155961–155970.

[Almatarneh et al. 2021] Almatarneh, S., Gamallo, P., ALshargabi, B., Al-Khassawneh, Y., and Alzubi, R. (2021). Comparing traditional machine learning methods for covid-19 fake news. In *2021 22nd International Arab Conference on Information Technology (ACIT)*, pages 1–4.

[Anoop et al. 2020] Anoop, K., Deepak, P., and V, L. L. (2020). Emotion cognizance improves health fake news identification. In *Proceedings of the 24th Symposium on International Database Engineering & Applications*, IDEAS '20, New York, NY, USA. Association for Computing Machinery.

[Bastos and Mercea 2019] Bastos, M. T. and Mercea, D. (2019). The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54.

[Boleda 2020] Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

[Cabral et al. 2021] Cabral, L., Monteiro, J. M. S., 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 *ICEIS*.

[Ciampaglia et al. 2015] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

[Cordeiro and Pinheiro 2019] Cordeiro, P. R. and Pinheiro, V. (2019). Um corpus de notícias falsas do twitter e verificação automática de rumores em lingua portuguesa. In *STIL-Brazilian Symposium in Information and Human Language Technology. IEEE, Salvaldor, BA, Brazil*, pages 220–228.

[Cruz et al. 2021] Cruz, R., Neto, G. N., and Anchiêta, R. (2021). Detecting misinformation in tweets related to covid-19. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 280–289, Porto Alegre, RS, Brasil. SBC.

[Dagan et al. 2005] Dagan, I., Glickman, O., and Magnini, B. (2005). The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

[Dahlgren 2018] Dahlgren, P. (2018). Media, knowledge and trust: The deepening epistemic crisis of democracy. *Javnost - The Public*, 25(1-2):20–27.

[Dantas and Deccache-Maia 2020] Dantas, L. F. S. and Deccache-Maia, E. (2020). Divulgação científica no combate às fake news em tempos de covid-19. *Research, Society and Development*, 9(7):e797974776–e797974776.

[Devlin et al. 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dryhurst et al. 2020] Dryhurst, S., Schneider, C. R., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., Spiegelhalter, D., and van der Linden, S. (2020). Risk perceptions of covid-19 around the world. *Journal of Risk Research*, pages 1–13.

[Dumitrache et al. 2018] Dumitrache, A., Aroyo, L., and Welty, C. (2018). Crowdsourcing semantic label propagation in relation classification. *CoRR*, abs/1809.00537.

[Evert 2010] Evert, S. (2010). Distributional semantic models. In *NAACL HLT 2010 Tutorial Abstracts*, pages 15–18, Los Angeles, California. Association for Computational Linguistics.

[Fonseca et al. 2016] Fonseca, E., Santos, L., Criscuolo, M., and Aluisio, S. (2016). Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

[Forelle et al. 2015] Forelle, M., Howard, P., Monroy-Hernández, A., and Savage, S. (2015). Political bots and the manipulation of public opinion in venezuela. *arXiv preprint arXiv:1507.07109*.

[Giordano et al. 2020] Giordano, G., Mottola, S., and Beatrice, P. (2020). A short review of some mathematical methods to detect fake news. International Journal of Circuits, Systems and Signal Processing, 14:255–265.

[Hartmann et al. 2017] Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

[Jacobsen and Vraga 2020] Jacobsen, K. H. and Vraga, E. K. (2020). Improving communication about covid-19 and emerging infectious diseases. *European journal of clinical investigation*, 50.

[Joachims 2002] Joachims, T. (2002). Learning to classify text using support vector machines - methods, theory and algorithms. In *The Kluwer international series in engineering and computer science*.

[Jones 2019] Jones, M. O. (2019). The gulf information war— propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. *International journal of communication*, 13:27.

[Krause et al. 2020] Krause, N. M., Freiling, I., Beets, B., and Brossard, D. (2020). Fact-checking as risk communication: the multi-layered risk of misinformation in times of covid-19. *Journal of Risk Research*, pages 1–8.

[Li et al. 2003] Li, Y., Bandar, Z., and Mclean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

[Lorena and Carvalho 2003] Lorena, A. C. and Carvalho, A. C. P. d. L. F. (2003). Introdução às máquinas de vetores suporte (support vector machines).

[Marín and Arroyo 2019] Marín, I. P. and Arroyo, D. (2019). Fake news detection. In *Computational Intelligence in Security for Information Systems Conference*, pages 229–238. Springer.

[Medeiros and Braga 2020] Medeiros, F. and Braga, R. (2020). Fake news detection in social media: A systematic review. In *Anais do XVI Simpósio Brasileiro de Sistemas de Informação*, Porto Alegre, RS, Brasil. SBC.

[Meleo-Erwin et al. 2017] Meleo-Erwin, Z., Basch, C., MacLean, S. A., Scheibner, C., and Cadorett, V. (2017). "to each his own": Discussions of vaccine decision-making in top parenting blogs. *Human vaccines & immunotherapeutics*, 13(8):1895–1901.

[Messeder Neto 2019] Messeder Neto, H. (2019). A divulgação científica em tempos de obscurantismo e de fake news: contribuições histórico-críticas. In Rocha, M. and Oliveira, R., editors, *Divulgação Científica: Textos E Contextos*. Livraria da Física, São Paulo, 1 edition.

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mitra et al. 2018] Mitra, B., Craswell, N., et al. (2018). *An introduction to neural information retrieval*. Now Foundations and Trends Boston, MA.

[Monteiro et al. 2018] Monteiro, R. A., Santos, R. L., Pardo, T. A., De Almeida, T. A., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese:

New corpus and automatic detection results. In *International Conference on Computational Processing of the Portuguese Language*, pages 324–334. Springer.

[Oshikawa et al. 2020] Oshikawa, R., Qian, J., and Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093.

[Paka et al. 2021] Paka, W. S., Bansal, R., Kaushik, A., Sengupta, S., and Chakraborty, T. (2021). Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393.

[Pawar et al. 2017] Pawar, S., Ramrakhiyani, N., Hingmire, S., and Palshikar, G. K. (2017). Topics and label propagation: Best of both worlds for weakly supervised text classification. *arXiv preprint arXiv: 1712.02767.*

[Pérez-Rosas et al. 2018] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics.

[Plous 1993] Plous, S. (1993). *The psychology of judgment and decision making.* Mcgraw-Hill Book Company.

[Priya and Kumar 2021] Priya, A. and Kumar, A. (2021). Deep ensemble approach for covid-19 fake news detection from social media. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 396–401.

[Ruediger 2017] Ruediger, M. A. (2017). Robôs, redes sociais e política no brasil: estudo sobre interferências ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018.

[Ruiz and Okano 2019] Ruiz, E. and Okano, E. (2019). Using linguistic cues to detect fake news on the brazilian portuguese parallel corpus fake. br. In *Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology*, pages 181–189.

[Sanh et al. 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[Schmidt et al. 2018] Schmidt, A. L., Zollo, F., Scala, A., Betsch, C., and Quattrociocchi, W. (2018). Polarization of the vaccination debate on facebook. *Vaccine*, 36(25):3606–3612.

[Shao et al. 2017] Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., and Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96:104.

[Shao et al. 2018] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

[Sharma et al. 2020] Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., and Liu, Y. (2020). Coronavirus on social media: Analyzing misinformation in twitter conversations. *CoRR*, abs/2003.12309.

[Silva et al. 2020] 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.

[Tandoc Jr et al. 2018] Tandoc Jr, E. C., Lim, Z. W., and Ling, R. (2018). Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

[Uscinski and Butler 2013] Uscinski, J. E. and Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, 25(2):162–180.

[van Dijck and Alinejad 2020] van Dijck, J. and Alinejad, D. (2020). Social media and trust in scientific expertise: Debating the covid-19 pandemic in the netherlands. *Social Media+ Society*, 6(4):2056305120981057.

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

[Vraga and Bode 2017] Vraga, E. K. and Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645.

[Wadden et al. 2020] Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

[Wang et al. 2011] Wang, B., Shen, Y., and Liu, Y. (2011). Integrating distance metric learning into label propagation model for multi-label image annotation. In *2011 18th IEEE International Conference on Image Processing*, pages 3649–3652.

[Wang et al. 2018] Wang, P., Angarita, R., and Renna, I. (2018). Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In *Companion Proceedings of the The Web Conference 2018*, pages 1557–1561.

[World Health Organization 2020] World Health Organization (2020). Novel coronavirus (2019-ncov) situation report - 13. Disponível em: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf.

[Yiannakoulias et al. 2019] Yiannakoulias, N., Slavik, C. E., and Chase, M. (2019). Expressions of pro-and anti-vaccine sentiment on youtube. *Vaccine*, 37(15):2057–2064.

[Zanzotto 2019] Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.

[Zhou et al. 2020] Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020). ReCOVery: A multimodal repository for covid-19 news credibility research. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.*

[Zhu and Ghahramani 2002] Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation.