Computational and Mathematical Organization Theory
https://doi.org/10.1007/s10588-022-09369-w

MANUSCRIPT

Fake or not? Automated detection of COVID-19 misinformation and disinformation in social networks and digital media

Izzat Alsmadi[1] · **Natalie Manaeva Rice**[2] · **Michael J. O'Brien**[3]

Abstract
With the continuous spread of the COVID-19 pandemic, misinformation poses serious threats and concerns. COVID-19-related misinformation integrates a mixture of health aspects along with news and political misinformation. This mixture complicates the ability to judge whether a claim related to COVID-19 is information, misinformation, or disinformation. With no standard terminology in information and disinformation, integrating different datasets and using existing classification models can be impractical. To deal with these issues, we aggregated several COVID-19 misinformation datasets and compared differences between learning models from individual datasets versus one that was aggregated. We also evaluated the impact of using several word- and sentence-embedding models and transformers on the performance of classification models. We observed that whereas word-embedding models showed improvements in all evaluated classification models, the improvement level varied among the different classifiers. Although our work was focused on COVID-19 misinformation detection, a similar approach can be applied to myriad other topics, such as the recent Russian invasion of Ukraine.

Keywords Coronavirus · COVID-19 · Disinformation · Learning models · Misinformation

✉ Izzat Alsmadi
izzat.alsmadi@tamusa.edu

[1] Department of Computing and Cyber Security, Texas A&M University–San Antonio, San Antonio, USA

[2] Center for Information and Communication Studies, University of Tennessee, Knoxville, USA

[3] Department of Communication, History, and Philosophy, Department of Life Sciences, Texas A&M University, San Antonio, USA

Published online: 25 November 2022

1 Introduction

The first cases of the SARS CoV-2 novel coronavirus (COVID-19) were reported in December 2019 in Wuhan, China. The World Health Organization (WHO) coronavirus dashboard (WHO 2022) reported that by March 9, 2022, there were almost 450 million confirmed cases across the globe and over 6 million deaths. For the United States, WHO reported over 78 million confirmed cases and over 950,000 deaths. In addition to the deaths and damage to public health, the COVID-19 pandemic unleashed major disruptions around the globe in terms of economy, education, and society in general (Alenezi and Alqenaei 2021).

Although there are other important factors that have contributed to the COVID-19 pandemic, there is strong consensus among researchers and public-health experts that the spread of COVID-19-related misinformation and disinformation on social- and digital-media platforms are major contributors (Tasnim et al. 2020; Roozenbeek et al. 2020; Vériter et al., 2020; Horawalavithana et al. 2021; Kricorian et al. 2021; Neely et al. 2022). As Tasnim et al. (2020:171) stressed, COVID-19 misinformation "is masking healthy behaviors and promoting erroneous practices that increase the spread of the virus and ultimately result in poor physical and mental health outcomes among individuals."

Leaders of various international organizations, including the United Nations and WHO, have called for special attention to be directed to the problem of misinformation and other types of falsehoods regarding the COVID-19 pandemic, calling it an "infodemic" (WHO 2021). WHO defined an infodemic as "too much information including false or misleading information in digital and physical environments during a disease outbreak," stressing that it can cause confusion and risk-taking behaviors that can harm public health.

Roozenbeek et al. (2020) found that although the beliefs in COVID-19 misinformation might not be prevalent in several countries, including the United Kingdom, the United States, Ireland, Mexico, and Spain, a substantial proportion of respondents in each country viewed COVID-19 misinformation as highly reliable. In the practical realm, that study also found that those respondents who believed in COVID-19 misinformation were also less likely to comply with public-health guidance.

1.1 COVID-19 disinformation and misinformation

In addition to the spread of misinformation, there have been numerous confirmed reports of disinformation campaigns directed and implemented by several state actors, including Russia, China, and Iran (Gradon 2020; Bright et al. 2020; Dubowitz and Ghasseminejad 2020; Hotez 2021; Horawalavithana et al. 2021). Although there are various ways to distinguish between misinformation and disinformation, we follow Jack (2017) in defining misinformation as information whose inaccuracy is unintentional and disinformation as information that is deliberately false or misleading (O'Brien and Alsmadi 2021). This distinction, based on intent, is also evident in the definitions provided by the Centers for Disease Control (CDC), which defines misinformation as "false information shared by people who do not intend to mislead

others" and disinformation as "false information deliberately created and disseminated with malicious intent" (CDC 2021).

In some cases, there is evidence that many disinformation campaigns are coordinated. For example, in 2020 European researchers confirmed that Russia and China were coordinating and synchronizing their efforts by producing similar messages and narratives and boosting and spreading each other's messages (Vériter et al. 2020). The topic of COVID-19 vaccinations has also been used in state-run information-warfare efforts, with researchers uncovering the efforts by the Russians to boost the popularity and sales of Russia-produced Sputnik V vaccine by spreading disinformation and undermining public trust in

vaccines produced in Western countries (Hotez 2021; Horawalavithana et al. 2021; U.S. Agency for Global Media 2021).

An additional difficulty in dealing with state-sponsored disinformation campaigns is that during ongoing efforts by state actors to influence democratic process or shape public opinion through social and digital media, those efforts seem to become more and more sophisticated and successful. For example, analysis of social-media posts produced by the Internet Research Agency, which represented a part of Russian efforts to influence U.S. elections, has shown that techniques and messages evolved over time in order to make them more effective (Ruck et al. 2019). Analysis of bot activity conducted by actors affiliated with the Russian government have also demonstrated that those techniques are evolving and becoming more sophisticated (Alsmadi and O'Brien 2020). This indicates that detecting and countering state-sponsored disinformation campaigns related to COVID-19 present additional challenges compared to other types of misinformation.

1.2 COVID-19 misinformation, disinformation, and vaccine hesitancy

Since the development and public rollout of COVID-19 vaccines, the spread of false information on social media and other digital platforms has been described by public-health officials and researchers as directly contributing to the high number of unvaccinated individuals in the United States and abroad (Dror et al. 2020; Kricorian 2021; Puri et al. 2020). Roozenbeek et al. (2020) found that those respondents who believed COVID-19 misinformation claims and questioned valid science-based claims were also less likely to get vaccinated or to recommend vaccinations to their friends and family, indicating a direct link between susceptibility to misinformation and the reduced likelihood of vaccination and adherence to health standards. A study of U.S. respondents by Kricorian et al. (2021) confirmed that misinformation about COVID-19 and vaccines was prevalent among those who refused to be vaccinated.

Further, analysis of survey data conducted by Neely et al. (2022) confirmed the previously detected link between misinformation and hesitancy among U.S. respondents. According to their analysis, although high levels of exposure to the misinformation were detected among the respondents, the exposure to false information was directly correlated with vaccine hesitancy, with politicization as a major contributing factor.

According to the New York Times COVID-19 Vaccination Tracker, by February 2022 63% of the global population had received at least one dose of the COVID-19

vaccine (New York Times 2022), whereas in the United States the level of partially vaccinated reached 75%; fully vaccinated 64%; and those who had received booster shots 27%. With the level of vaccination in the United States and around the globe being lower than what is needed to achieve herd immunity, proliferation of misinformation and disinformation remains "a significant impediment to the attainment of herd immunity and the end of the COVID-19 pandemic" (Neely et al. 2021:179).

Since the beginning of the pandemic, there has been a growing number of calls by experts and stakeholders to monitor and combat the spread of COVID-19 and vaccination-related misinformation and disinformation on social media, digital media, and other platforms. U.N. Secretary General Antonio Guterres called for additional efforts to stop the spread of false information and conspiracy theories that have a direct negative impact on efforts to curtail the pandemic, directly addressing social-media companies to flag and remove harmful content and to "remove racist, misogynist and other harmful content" (CBS News 2020). WHO partnered with a number of major tech companies, including Facebook, Twitter, and YouTube, to detect and delete COVID-19-related misinformation and promote legitimate updates from healthcare organizations (Statt 2020). However, recent results showed that the tech companies often fail to adequately monitor COVID-19 falsehood and delete such content in a timely manner (Brindha et al. 2020; Wardle and Singerman 2021).

1.3 COVID-19 and machine-learning efforts

As a result of the high volume of misinformation, disinformation, and other types of falsehoods on various social- and digital-media platforms, automated detection of such false or inaccurate information has recently gained importance and has become a primary detection technique (Tacchini et al. 2017; Thota et al. 2018; Ruchansky et al. 2017). Several recent studies have highlighted the need to apply machine learning and other techniques to the problem of widespread COVID-19 misinformation and disinformation. For example, Tasnim et al. (2020) have called for using advanced technology such as natural language processing (NLP) or data-mining approaches to detect and remove misinformation and other types of falsehoods with no basis in science from digital platforms.

Employment of advanced technologies to detect misinformation from social media and other digital sources is a robust and developing field that produces successful results (Shu et al. 2017). Although algorithms have been successfully employed to identify false information, this work has its own set of unique challenges (Shu et al. 2017). Tasnim (2020), as well as Alenezi and Alqenaei (2021), argued that despite challenges, application of the same principles behind identifying and removing false COVID-19 information is both feasible and highly desirable. While research showed that misinformation spreads faster on social-media platforms than information from legitimate news sources (Tasnim et al. 2020), applying machine learning to detection of falsehoods might be a major tool in fighting the global pandemic.

Several studies have focused on developing tools for automatic detection of COVID-19-related misinformation using NLP approaches, including detection-of-misinformation videos on YouTube by leveraging user comments (Serrano et al.

2020) and classification of social-media posts containing misinformation based on health risks associated with them (Dharawat et al. 2020).

Hossain et al. (2020) pointed out that the existing misinformation-detection datasets were not effective for evaluating systems designed to detect COVID-19 misinformation resulting from the use of novel language and rapid changes in information. They also released the COVIDLIES1 dataset and evaluated existing NLP systems on that dataset.

Alenezi and Alqenaei (2021) proposed building machine-learning misinformation-detection models that target COVID-19 misinformation in social media. Specifically, they tested three detection models—long short-term memory (LSTM) networks, a multichannel convolutional neural network (MC-CNN), and k-nearest neighbors (kNN) on Twitter data and obtained results superior to those from previous studies.

The Bidirectional Encoder Representations from Transformers (BERT) language-representation model initially developed by Devlin et al. (2018) was successfully used by several research teams —for example, Wani et al. (2021) and Glazkova et al. (2021)—to evaluate social-media posts related to COVID-19 misinformation. Hamid et al. (2020) examined both falsehoods and 5G conspiracy theories, and Wahle et al. (2021) focused on using transformer-based models on five COVID-19 misinformation datasets that included a variety of sources such as social-media posts, news articles, and scientific papers.

For this paper, we evaluated the impact of using word- and sentence-embedding models and transformers on classification models. Our effort was motivated by several recent publications that indicate the advantage of using embedding models in general and sentence transformers (e.g., BERT) to improve the prediction of classification models (e.g., Ling et al. 2017; Liu et al. 2019; Hao et al. 2019; Ruas et al. 2020). Another motivation for using embedding models is related to our integration of different datasets related to COVID-19. We believe that integrating text from different datasets, while simultaneously employing word-embedding models, can help generalize results from classification models and help reduce possible bias in their predictions. The complexity of using several embedding models is related to the amount of time and resources needed to pre-train large datasets in each one of those models. The pre-trained models cannot be reused from one embedding model to another. However, for the same embedding model, pre-trained models can be used to pre-train and test new data.

2 Research questions and methods

The following two questions guided our research:

1. Why and how to integrate different text-based datasets? In the next section, research methods, we discuss some of the reasons for integrating different text-based datasets from the same domain. Here, the research question is related to how best to integrate datasets if they have different or somewhat different target-column labels. Although most COVID-19 misinformation datasets have a binary target column identifying whether the information is fake or not, we noticed that

the way they identify and describe what is fake and what is not fake is different. Additionally, the term misinformation has different meanings and interpretations.

2. Which embedding types or models improve COVID-19 misinformation-classification models? We evaluated several models—W2V, Glove, Google, Paragram, Wiki, and BERT—that are available for public use in research on COVID-19 integrated dataset. We observed similar goals in the use of those embedding models and our dataset's integration in producing pre-trained models that can be used beyond a single or a particular dataset.

Figures 1 and 2 summarize the two experimental tracks we followed. The first track focused on data-analytics models per the different individual datasets, whereas the second track focused on first combining those datasets before performing data-analytics tasks.

Feature extractions in text-based datasets are based primarily on the text column, and so we believe that text datasets in the same domain should be combined to improve the quality of classification models.

We think there is value in combining datasets:

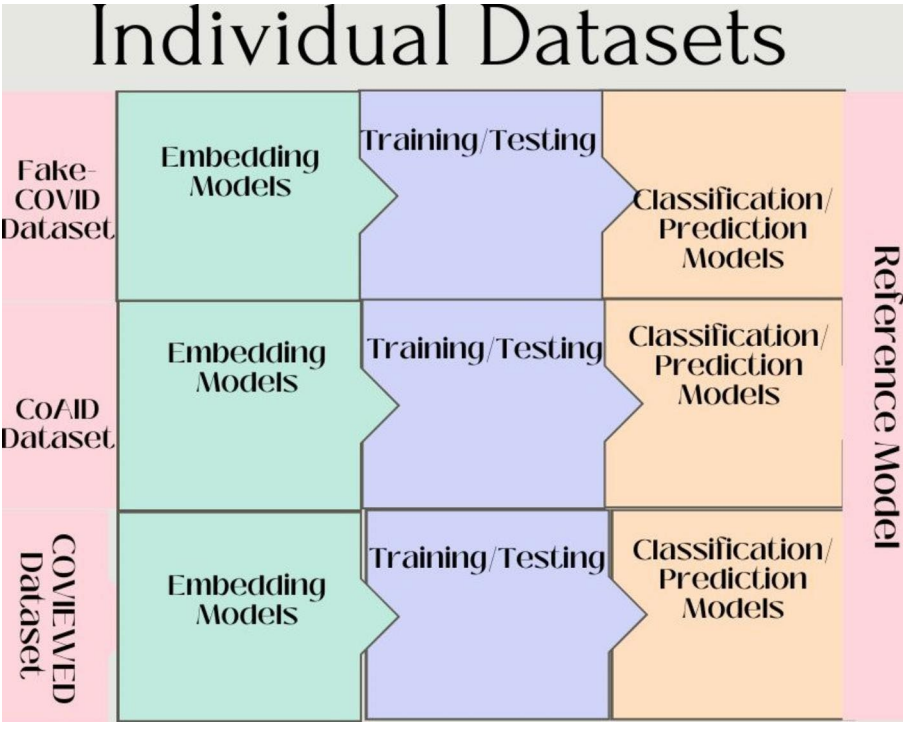

Fig. 1 Research methods: Individual datasets

1. Text-based datasets have similar feature-based extractions, i.e., from the single-text column, unlike nontext-based datasets, where features can be very different from one dataset to another.
2. In the same or similar domains, corpus-based features are expected to be similar. In other words, for machine-learning models, for the same domain and target, popular text-based features should be similar.
3. Reduces bias in input datasets. Here, bias refers to machine-learning bias, particularly when models perform well because of narrow or certain scopes. There are several aspects and reasons for bias in machine-learning models. One aspect of possible bias is related to the input dataset. Integrating different datasets for the same domain is expected to reduce such bias.

We began the experiments with basic models on the individual datasets. We then reused those models in the integrated datasets and introduced new models applied on the integrated dataset.

As implied in Fig. 2, our goal is to evaluate producing a reference model that can be used for other datasets beyond those used here.

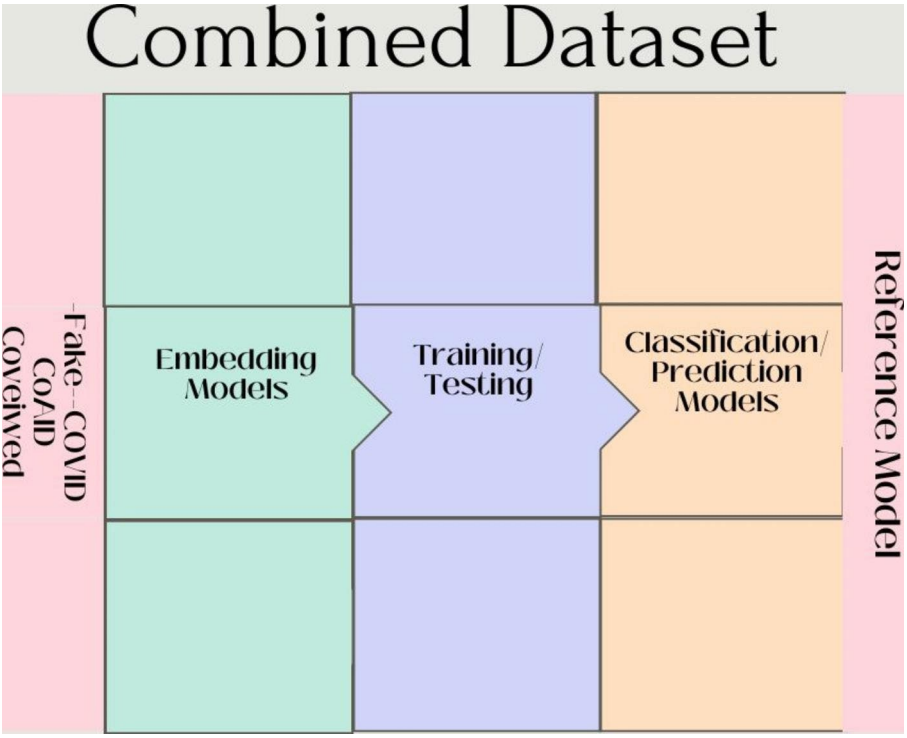

Fig. 2 Research methods: Combined dataset

3 Data, experiments, and analysis

We used three public datasets related to COVID-19 misinformation:

1. FakeCOVID (Shahi and Nandini 2020). The FakeCOVID dataset is multilingual, with 7623 news articles related to COVID-19. In addition to not fake/fake[1], some articles are labeled with partially not fake or partially fake. We used 6286 of those articles that had either fake or not-fake labels. The dataset is imbalanced; only 34 articles were labeled as not fake, and the rest were labeled as fake.
2. CoAID (COVID-19 heAlthcare mIsinformation Dataset) (Cui and Lee 2020). CoAID is a diverse COVID-19 healthcare-misinformation dataset that includes fake news from websites and social networks and also users' responses to that news. It includes 3,235 news stories, 294,692 related user engagements, 851 social-platform posts about COVID19, and ground-truth labels.
3. The COVIEWED project is an effort to combat COVID-19 misinformation. By initiating the effort through a public website (COVIEWED 2020), the project invites data-science researchers to present their ideas for achieving that goal. We used one subset from COVIEWED submissions, which includes a known list of COVID-19 misinformation collected from IDeas (2020). This list is labeled as fake claims. However, the definition of not fake claims was more generic, and the dataset includes many posts or comments from different websites that are citing COVID-19.

We noticed that the datasets have different interpretations of what is fake and what is not fake. Although we acknowledge that classifiers' accuracy would be impacted with such integration, we nonetheless believe that this combination of datasets will achieve two main goals:

1. To produce models that generalize to multiple disinformation topics.
2. To reduce possible bias in proposed models. Bias can exist or be introduced to machine-learning models in several different ways. For example, classification models that are generated based on specific datasets can be biased or overfitted as a result of issues in the input data (e.g., the features, how data are collected, and how target columns are interpreted). They might work well in the evaluated datasets but poorly in any other dataset.

3.1 Results from the FakeCOVID dataset

We initially evaluated three single classifiers: logistic regression (LR), support vector classifier (SVC), and naive Bayes (NB). As expected, because of the imbalance in the dataset, one class label that had a relatively large number of instances reported very good accuracy whereas the other did not. Table 1 summarizes the performance metrics for all three classifiers.

To summarize:

Table 1 Classification metrics for the FakeCOVID dataset

Classifier	Fake Claims			Not Fake Claims			Accuracy
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
LR (CV)	1	1	1	1	0.4	0.57	1
LR (TFIDF)	1	1	1	0	0	0	1
SVC (CV/ TFIDF)	1	1	1	0	0	0	1
NB (CV/ TFIDF)	1	1	1	0	0	0	1

- Performance metrics for fake claims are very high, as there are enough data to help machine-learning algorithms learn and predict.
- On the other hand, those metrics are very low for the not-fake claims, as there are not enough data to build a reliable classification model.
- Unlike the SVC and NB classifiers, which reported the same results for both CV and TFIDF terms/features' extractions, the LR classifier produced different results for the two approaches.
- Overall accuracy is high, as the majority of the dataset is on the fake-claims side. In comparison with the original paper that produced this dataset (Shahi and Nadini 2020), we found better results in terms of all reported metrics, accuracy, precision, recall, and F-score.

3.2 Results from the CoAID dataset

Using the same three single classifiers—LR, SVC, and NB—we found that the CoAID dataset showed variation in performance results (Table 2). The sample dataset had balanced numbers from both label types. All three classifiers showed different results between CV and TFIDF, with CV producing better accuracy in all three classifiers.

3.3 Results from the COVIEWED dataset

The COVIEWED dataset had more not-fake claims than fake claims, and thus performance metrics are high on not-fake claims and very low on fake claims. Accuracy is similar for all classifier models (Table 3).

Table 2 Classification metrics for CoAID dataset

Classifier	Fake Claims			Not Fake Claims			Accuracy
	Precision	Recall	F1-score	Recall	Precision	F1-score	
LR (CV)	0.88	0.53	0.66	0.87	0.98	0.92	0.87
LR (TFIDF)	0.90	0.37	0.52	0.83	0.99	0.99	0.84
SVC (CV)	0.95	0.53	0.68	0.87	0.99	0.93	0.88
SVC (TFIDF)	0.95	0.46	0.62	0.85	0.99	0.92	0.86
NB (CV)	0.88	0.71	0.78	0.91	0.97	0.94	0.91
NB (TFIDF)	1	0.37	0.54	0.83	1.00	0.91	0.85

Table 3 Classification metrics for COVIEWED dataset

Classifier	Fake Claims			Not Fake Claims			Accuracy
	Precision	Recall	F1-score	Recall	Precision	F1-score	
LR (CV)	0.24	0.05	0.08	0.92	0.99	0.95	0.91
LR (TFIDF)	0.00	0.00	0.00	0.92	1.00	0.96	0.92
SVC (CV)	1.00	0.00	0.01	0.92	1.00	0.96	0.92
SVC (TFIDF)	0.75	0.01	0.02	0.92	1.00	0.96	0.92
NB (CV)	0.17	0.01	0.02	0.92	0.99	0.95	0.91
NB (TFIDF)	0.00	0.00	0.00	0.92	1.00	0.96	0.92

3.4 Combination of COVID-19 misinformation datasets

As mentioned earlier, datasets related to misinformation are inconsistent in terms of how they label information/misinformation. As a result, it is difficult to (1) integrate different datasets with each other and/or (2) transfer models and knowledge from one dataset to another. Our goal here is to show different ways of dealing with such issues.

We focused on combining only two columns, the text column and the label column, and ignored all other columns that can be different among datasets.

We generalized the terminology among the different datasets. Our combined dataset included a more generic binary label of fake versus not fake to accommodate less-generic labels used by the different datasets. For example, a true claim indicates telling a correct story but not necessarily with correct information. In our combined label, this is "not fake." We could extend this approach by combining misinformation datasets from different categories (e.g., false claims, fake news, hoaxes, spam, and insincere questions) and creating a broad binary label, such as fake/not fake and yes/no.

We aggregated data about COVID-19 misinformation from different sources for several reasons, one related to bias and overfitting issues. Bias refers to models that can be highly accurate in terms of performance metrics but which represent only a subset of reality due to their focus on some data points while ignoring others. Overfitting in data analytics refers to a problem when models work well in one dataset or a subset of a dataset but poorly when applied to different datasets that were not part of model learning or testing.

We aggregated the three datasets and used instances from all three in both training and testing. The combined dataset has a total of 20,563 claims, 7,905 of which are fake claims and 12,658 of which are not-fake claims. Preliminary text analysis for the fake versus not-fake showed some differences. The first feature we evaluated was the word count.

We created word clouds for fake versus not-fake combined text, as shown in Fig. 3. Word clouds show the top words in each group and highlight the different focuses between fake/not fake discussions.

We can summarize top words between the two word clouds as

1. Top words in both clouds: COVID-19, coronavirus, virus, people, hospital, say/said, novel, China;

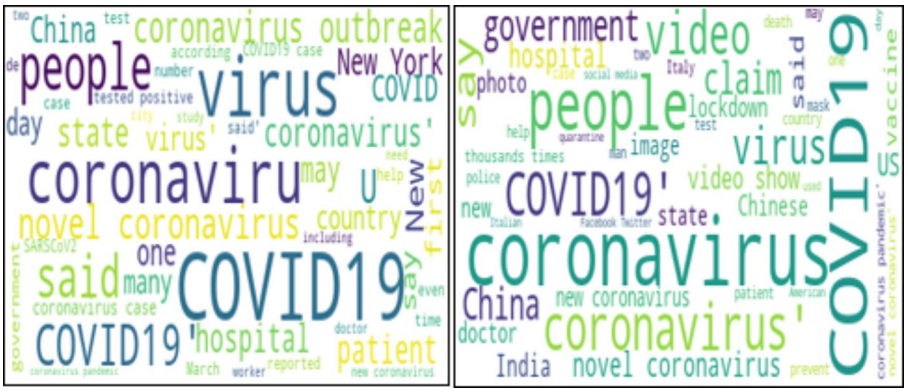

Fig. 3 Word cloud for the title column (fake text, right, versus not fake, left)

2. Top words in not-fake content: outbreak, patient, New, New York; and
3. Top words in fake content: lockdown, vaccine, India, Chinese, video.

Then we did the same using the content column, as shown in Fig. 4.

4 Features and classification models

One important step in text analysis is to evaluate features that can produce classification or prediction models with high accuracy. We evaluated two popular approaches—count vectors (CV) and term frequency/inverse document frequency (TF/IDF). Figure 5, left, shows an assessment of using CV for several classifiers. For our four evaluated performance metrics, Precision, Recall, F1-Score, and Accuracy, except for KNN, most classifiers had values between 70% and 80% with all metrics. Again, except for KNN, all classifiers showed similar values in metrics between fake and not-fake claims. Figure 5, right, shows similar results for TF/IDF.

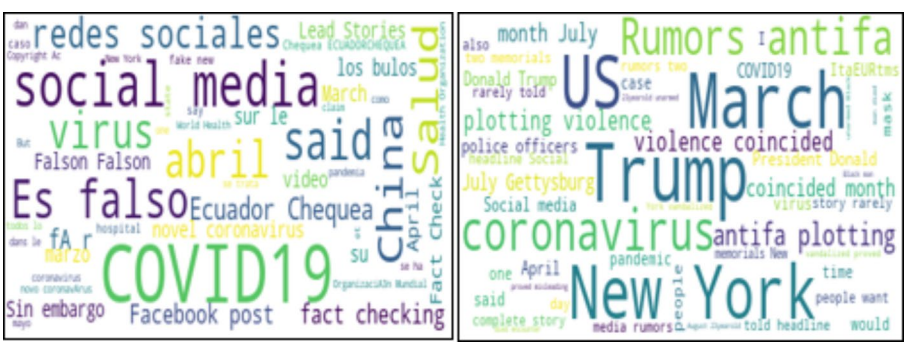

Fig. 4 Word cloud for the content column (fake text, right, versus not fake, left)

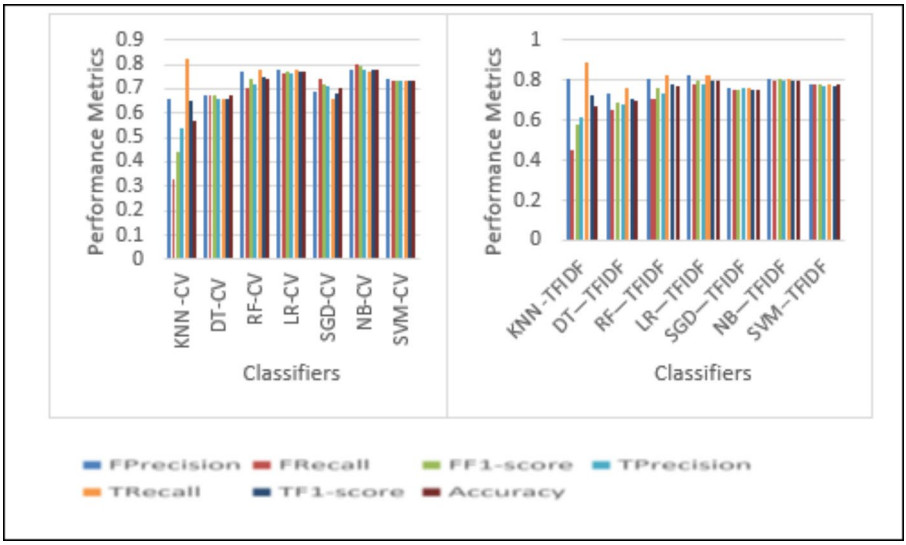

Fig. 5 Classifiers' performance metrics: (CV, left, TF/IDF, right), 100 terms, no embedding

The previous experiments used an initial fixed set of terms/features (100). Our next goal was to evaluate the impact of increasing the number of input terms/features on the performance of classification models. We also wanted to assess the impact of increasing the input dataset from 2000 to 15,000 claims while ensuring that the same number of fake and not fake claims was used in each model. Each classifier had a few input variables. As previous results showed similar results between CV and TF/IDF, and to reduce redundancy, we report results from only one model, CV text-feature extraction. Table 4 summarizes results from evaluating the number of terms on classifiers' performance metrics.

The table shows that of all evaluated classifiers, the Decision Tree (DT) shows the lowest accuracy in both evaluated settings. Additionally, the classifier showed insensitivity to increasing the number of terms in the module. Its best performance metrics were achieved with a relatively small number of terms. Adding more terms did not improve performance metrics but rather had the opposite results in some cases.

Table 4 Classification accuracy versus the number of model input terms

Type	Terms	PF	RF	F1F	PT	RT	F1T	Acc
DT1	6000	0.78	0.76	0.77	0.77	0.79	0.78	0.78
LGR1	50,000	0.88	0.79	0.83	0.81	0.90	0.85	0.84
SGD1	50,000	0.89	0.79	0.84	0.81	0.90	0.86	0.85
SVC1	50,000	0.87	0.80	0.83	0.82	0.88	0.85	0.84
DT2	1000	0.78	0.75	0.76	0.76	0.79	0.77	0.77
LGR2	50,000	0.88	0.80	0.84	0.82	0.90	0.86	0.85
SGD2	50,000	0.89	0.81	0.85	0.83	0.90	0.86	0.86
SVC3	50,000	0.87	0.82	0.84	0.83	0.88	0.85	0.85

All other evaluated classifiers showed sensitivity to increasing the number of terms as input features to the classification model. As a cost, increasing the number of terms will increase the model complexity and impact its efficiency.

5 Learning with word- and sentence-embedding models

Word- and sentence embedding is a method for obtaining a context-dependent vectorized representation for every word/sentence in a text corpus. This representation allows comparison of words in embedding space: Words spaced closer together have a similar meaning and/or connotation, whereas words far apart are very dissimilar. Word-embedding data are existing, pre-trained distributed word representations. The main task is to determine the most qualitative word embeddings. In the process, distributional models are generated over different sources such as Wikinews, news articles, Google News, and BERT.

Recent state-of-the-art word-embedding models such as BERT have proven successful in obtaining relevant word embeddings for practical applications such as language translation. All terms in the corpus are embedded. The BERT sentence-transformers repository allows training and transformer models to generate sentence and text embeddings (Reimers 2019). Sentence BERT uses a Siamese network-like architecture to provide two sentences as an input. The sentences are then passed to BERT models and a pooling layer to generate their embeddings (Huilgol 2020).

To evaluate the impact of using word embeddings, we used the same classification settings of the previous experiment with the addition of using word embeddings. Before using the training and testing data from COVID-19 claims, both were trained with the BERT embedding model. The trained outputs were used as input for all classifiers. Figure 6 shows a summary of the accuracy metric for all classifiers. Except for Decision Tree models, classification models showed improvement in all performance metrics when using embedding models. Unlike in previous experiments, no classifier showed sensitivity to an increase in the model's number of terms.

5.1 Most-informative features

Classification and prediction models are based on input features. In text analytics, those features can be extracted from either text statistics or text corpus. In text corpus, the default approach is to use tokens—words, phrases, n-grams, and the like—as features. The process starts with all text. Different preprocessing steps such as stopwords removals and stemming can be applied to produce a preprocessed corpus. The analysis then focuses on producing the most-informative features that can predict the classification target class. Below we present three examples of the approaches we evaluated to extract the most-informative single-word features. Due to space limitation, we show results from only one experiment, TFIDF most-informative features (Table 5).

Looking at the most-informative text-based terms, we can see that they may reveal more about the particularities and properties of the datasets used rather than any objective truth about which words are good indicators of fake news. A large majority

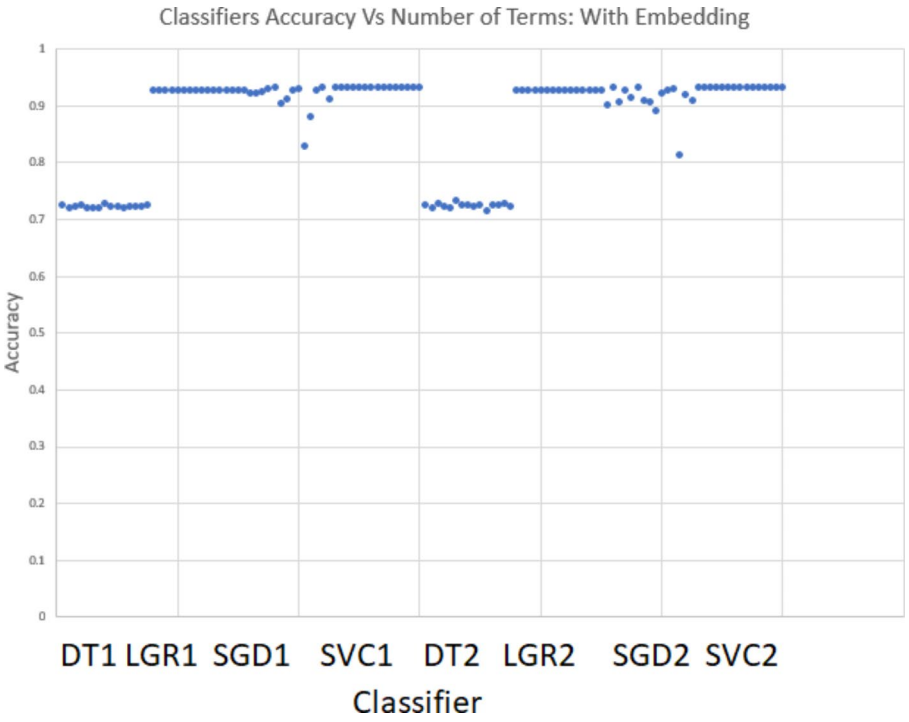

Fig. 6 Classifiers' accuracy versus the number of terms, with embedding

of the top terms (not shown in the previous tables) are simply words that point to a specific domain or publisher. Similar to findings mentioned in other research (e.g., Fairbanks et al. 2018), informative terms in the not fake-claims section include those that typically exist in news articles, whereas informative terms in the fake-claims section include highly specialized terms, indicating that they refer to specific conspiracy theories. Table 6 shows a sample from another approach that integrates the logistic-regression (LR) classifier with the chi-square feature-selection method. Chi-square values show the significance of the term on LR classification or on making a prediction of an instance target label.

5.2 Comparison of word-embedding models

We used several word-embedding models under the same experimental settings to extend our assessment of using word-embedding models in COVID-19 fake-news detection. The specific word-embedding models we used were W2V, Glove, Google, Paragram, Wiki, and BERT. Overall, SGD and logistic-regression classifiers scored the highest accuracy of values—between 86% and 87% in most embedding models—whereas the MLP Classifier scored the lowest in most experiments (Fig. 7).

Table 5 Most informative features: TFIDF.

FAKE		NOT FAKE	
exploring	backed	automated	stem
clampdown	proves	SARS	shortness
Gates	Ghana	lessons	regions
encourage	Leganés	Asian	March
Woolfson	Brazilian	lives	material
useful	penetration	huge	overcrowded
claiming	conclusion	residents	COVID
image	video	key	Latin
foregoing	proving	webMD	tweets
China	inclusion	monkeys	Quebec
airplane	technique	flouting	January
catching	allowed	receives	admin
prevents	ivermectin	critical	week
seconds	sickness	reactivating	overwhelmed
homemade	antibiotics	handled	Davis
curfew	fibrous	waiting	protective
exacerbated	demonstrated	provide	clinicians
Caixin	proven	meant	changer
hypoxia	APnews	rebate	plans
UNICEF	garden	guidelines	diplomat
mothers	antibiotic	resources	normal
photo	Draco	skepticism	rapid
disappointing	dengue	calculate	continues
leaked	harmful	commentary	briefing
analgetics	dampening	build	count

Table 6 Top terms using LR and chi-square

Term	Chi Square	Term	Chi Square
video	100.15	longside	22.53
virus	68.43	breath	22.23
Facebook	65.20	Paulo	22.15
shared	62.49	claim	21.62
posts	56.01	photo	21.56
lockdown	45.15	streets	21.36
shows	42.04	India	21.23
said	40.27	kills	21.13
people	39.53	salt	21.00
water	33.62	Brazilian	19.20
will	33.31	quarantine	18.71
cure	30.56	cures	18.62
photo	27.94	gargling	18.00
image	25.97	Indian	17.75
drinking	24.14	warm	16.20
Twitter	24.00	kill	16.07
lemon	23.00	president	16.00

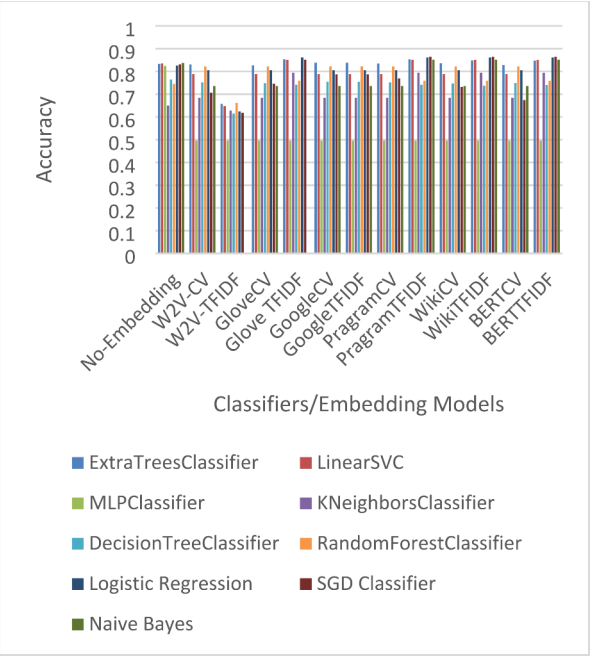

Fig. 7 Classifiers versus word-embedding models

6 Conclusion

As Tasnim et al. (2020) stated, providing a variety of stakeholders, including social-media companies, healthcare professionals, mass media, and other actors, with the latest results of research related to battling the spread of misinformation, disinformation, and other untrustworthy online information related to COVID-19 is an important step in bringing closer the end of the devastating global pandemic. Given that previous studies have demonstrated a direct link between COVID-19 misinformation and an unwillingness to follow public-health measures, effective application of machine-learning techniques to detect misinformation and disinformation in social and digital platforms is becoming an increasingly important tool in the global fight against the deadly disease.

We evaluated some of the challenges related to using some public-misinformation datasets to extract relevant knowledge. Misinformation these days refers to a spectrum of terminologies and concepts that can differ from each other in many respects. As a result, analytic models that are produced based on those datasets can be biased and may not work well with different datasets. We combined several misinformation-related datasets that discuss different aspects of misinformation related to COVID-19. As the three datasets we used have imbalanced class labels, one advantage of the integration was fixing such imbalance.

The combination process was simple, as text-based datasets focused on two main columns, text and label. The major challenge in the integration will be when class labels differ among datasets. With misinformation datasets, we found the best approach is to broadly categorize all misinformation labels under one category and

similarly combine the opposite class labels. This can make models more general and less biased, although prediction accuracy may be impacted.

We focused our analysis of how some recent text analyses featuring extraction and prediction techniques can impact prediction models' performance. We observed that some classifiers are more sensitive than others to the volume of search terms. We also observed that whereas word-embedding methods showed improvements in all evaluated classification models, the improvement level can vary among the different classifiers. Compared to word- and sentence-embedding models, our experiments showed that recent sentence transformers such as BERT showed better improvements on most classifiers.

For machine-learning models and tools to be accurate in terms of classification/prediction and to be transferable from one model or dataset to another, there is a need for common and unified terminologies for both information and misinformation. Again, although our work was focused on COVID-19-related information, a similar approach of combining datasets to improve performance in learning models could be used with a variety of other topics that are prone to high saturation of misinformation and disinformation, including political messages on social and digital media. For example, recent Russian aggression against Ukraine that employs a variety of information-warfare techniques targeting multiple audiences outside of Russia shows a growing need for using machine-learning techniques to identify such disinformation campaigns.

Note [1] For consistency, we used the terms "fake" and "not fake," whereas other authors or datasets sometimes use terms such as "true" and "false."

Acknowledgments We thank three anonymous reviewers for detailed suggestions on how to greatly improve the manuscript and Kathleen Carley for her advice and editorial help.

References

Alenezi MN, Alqenaei ZM (2021) Machine learning in detecting COVID-19 misinformation on Twitter. Future Internet 13(10):244. https://doi.org/10.3390/fi13100244

Bright J et al (2020) Coronavirus coverage by state-backed English-language news sources. Computational Propaganda Project, Oxford, Data Memo

Brindha D, Jayaseelan R, Kadeswaran S (2020) Social media reigned by information or misinformation about COVID-19: a phenomenological study. Social Sciences & Humanities Open. https://doi.org/10.2139/ssrn.3596058

CBC News (2021) COVID-19 pandemic unleashing 'tsunami of hate,' says UN chief. https://www.cbc.ca/news/world/coronvirus-un-fear-xenophobia-1.5561069

COVIEWED (2020) https://www.kaggle.com/trtmio/project-coviewed-subreddit-coronavirus-news-corpus

Cui L, Lee D (2020) CoAID: COVID-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885

Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dharawat A, Lourentzou I, Morales A, Zhai C (2020) Drink bleach or do what now? COVID-HeRA: a dataset for risk-informed health decision making in the presence of COVID19 misinformation. arXiv:2010.08743v1.

Dror AA, Eisenbach N, Taiber S, Morozov NG, Mizrachi M, Zigron A, Srouji S, Sela E (2020) Vaccine hesitancy: the next challenge in the fight against COVID-19. Euro J Epidemiol 35(8):775–779. https://doi.org/10.1007/s10654-020-00671-y
Dubowitz M, Ghasseminejad S (2020) Iran's COVID-19 disinformation campaign. Combatting Terrorism Center 13(6):40–48
Fairbanks J, Fitch N, Knauf N, Briscoe E (2018) Credibility assessment in the news: do we need to read. Proc of the MIS2 Workshop held in conjuction with 11th Intl Conf on Web Search and Data Mining February (pp. 799–800)
Glazkova A, Glazkov M, Trifonov T (2021) Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. arXiv:2012.11967
Gradoń K (2020) Crime in the time of the plague: fake news pandemic and the challenges to law-enforcement and intelligence community. Soc Register 4(2):133–148. https://doi.org/10.14746/sr.2020.4.2.10
Hamid A et al (2020) Fake news detection in social media using graph neural networks and NLP techniques: A COVID-19 use-case. arXiv preprint arXiv:2012.07517
Hao Y, Dong L, Wei F, Xu K (2019) Visualizing and understanding the effectiveness of BERT. https://doi.org/10.18653/v1/D19-142
Horawalavithana S, Silva RD, Nabeel M, Elvitigala C, Wijesekara P, Iamnitchi A (2021) Malicious and low credibility URLs on Twitter during the AstraZeneca COVID-19 vaccine development. https://doi.org/10.1007/978-3-030-80387-2_1
Hossain T, Logan IVRL, Ugarte A, Matsubara Y, Young S, Singh S (2020) COVIDLies: Detecting COVID-19 misinformation on social media. https://doi.org/10.18653/v1/2020.nlpcovid19-2.11
Hotez PJ (2021) Anti-science kills: from Soviet embrace of pseudoscience to accelerated attacks on US biomedicine. PLOS Bio 19(1). https://doi.org/10.1371/journal.pbio.3001068
Huilgol P (2020) Sentence embedding techniques one should know with Python codes. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/08/top-4-sentence-embedding-techniques-using-python/
IDeaS, Center for Informed Democracy & Social–cybersecurity (IDeaS). Carnegie Mellon University. https://www.cmu.edu/ideas-social-cybersecurity/research/coronavirus.html
Kricorian K, Civen R, Equils O (2021) COVID-19 vaccine hesitancy: misinformation and perceptions of vaccine safety. Hum Vac & Immuno 8:1–8. https://doi.org/10.1080/21645515.2021.1950504
Ling Y et al (2017) Integrating extra knowledge into word embedding models for biomedical NLP tasks. https://doi.org/10.1109/IJCNN.2017.7965957
Liu Y et al (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692
Neely SR, Eldredge C, Ersing R, Remington C (2021) Vaccine hesitancy and exposure to misinformation: a survey analysis. J Gen Intern Med 37(1):179–187
New York Times (2022) COVID Vaccination Tracker https://www.nytimes.com/interactive/2021/world/COVID-vaccinations-tracker.html
O'Brien M, Alsmadi I (2021) Misinformation, disinformation and hoaxes: what's the difference? Conversation. https://theconversation.com/misinformation-disinformation-and-hoaxes-whats-the-difference-158491
Puri N, Coomes EA, Haghbayan H, Gunaratne K (2020) Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vacc Immunother 16(11):2586–2593. https://doi.org/10.1080/21645515.2020.1780846
Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084. https://doi.org/10.18653/v1/D19-1410
Roozenbeek J, Schneider CR, Dryhurst S, Kerr J, Freeman AL, Recchia G, Van Der Bles AM, Van Der Linden S (2020) Susceptibility to misinformation about COVID-19 around the world. Roy Soc Open Sci 14(10):201199. https://doi.org/10.1098/rsos.201199
Ruas D et al (2020) Enhanced word embeddings using multi-semantic representation through lexical chains. Info Sci 532:16–32. https://doi.org/10.1016/j.ins.2020.04.048
Ruchansky N, Seo S, Liu Y (2017) CSI: a hybrid deep model for fake news detection. https://doi.org/10.1145/3132847.3132877
Ruck DJ, Rice NM, Borycz J, Bentley RA (2019) Internet Research Agency Twitter activity predicted 2016 US election polls. First Monday24(7)
Serrano JC, Papakyriakopoulos O, Hegelich S (2020) NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. https://aclanthology.org/2020.nlpcovid19-acl.17
Shahi GK, Nandini D (2020) FakeCOVID—A multilingual cross-domain fact check news dataset for COVID-19. arXiv preprint arXiv:2006.11343

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explorations Newsletter 1;19(1):22–36. https://doi.org/10.1145/3137597.3137600

Statt N (2020) Major tech platforms say they're 'jointly combating fraud and misinformation'about COVID-19. Verge. https://www.theverge.com/2020/3/16/21182726/coronavirus-covid-19-facebook-google-twitter-youtube-joint-effort-misinformation-fraud

Tacchini E et al (2017) Some like it hoax: automated fake news detection in social networks. arXiv:1704.07506.

Tasnim S, Hossain MM, Mazumder H (2020) Impact of rumors and misinformation on COVID-19 in social media. J Prev Med Public Health 53(3):171–174. https://doi.org/10.3961/jpmph.20.094

Thota A, Tilak P, Ahluwalia S, Lohia N (2018) Fake news detection: a deep learning approach. SMU Data Sci Rev 1(3):10

U.S. Agency for Global Media (2021) https://www.usagm.gov/2021/06/15/usagm-networks-investigate-russian-disinformation-about-western-vaccines/

Vériter SL, Bjola C, Koops JA (2020) Tackling COVID-19 disinformation: internal and external challenges for the European Union. Hague J Diplomacy 15(4):569–582

Wahle JP et al (2021) Testing the generalization of neural language models for COVID-19 misinformation detection. https://doi.org/10.1007/978-3-030-96957-8_33

Wani A, Joshi I, Khandve S, Wagh V, Joshi R (2021) Evaluating deep learning approaches for COVID19 fake news detection. https://doi.org/10.1007/978-3-030-73696-5_15

Wardle C, Singerman E (2021) Too little, too late: social media companies' failure to tackle vaccine misinformation poses a real threat. https://doi.org/10.1136/bmj.n26

World Health Organization (2021) Cross-regional statement on "infodemic" in the context of COVID-19. https://onu.delegfrance.org/IMG/pdf/cross-regional_statement_on_infodemic_final_with_all_endorsements.pdf