Fake or Not? Covid-19 Misinformation, Between Science and Social Networks*

Izzat Alsmadi[1][0000−0001−7832−5081] and Michael J. O'Brien[2]

[1] Texas A&M, San Antonio, ialsmadi@tamusa.edu
[2] Texas A&M, San Antonio, Mike.Obrien@tamusa.edu

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 fake or not. Eventually, this also impacts the ability to reuse knowledge and may bias machine learning models and their ability to judge whether a claim is false or not. In an effort to deal with these issues, we aggregated several COVID-19 misinformation datasets and compared differences between learning models from individual datasets versus the aggregated ones.

Keywords: Coronavirus · COVID-19 · Fake news · Learning models · Misinformation.

1 INTRODUCTION

The COVID-19 pandemic created a worldwide impact on all human-life aspects. The pandemic exploded beyond health and life concerns. People worldwide were forced to change their behaviors, especially on how they communicate and interact with one another. The COVID pandemic has some unique aspects that help the spread of misinformation. People have no clear or credible information about COVID-19, its origin, possible treatments, and the like. People exchange lots of information and misinformation about COVID-19, with a focus on subjects such as:

- The origin of COVID-19: The main story of the birth of COVID-19 is related to patients in Wuhan, China, with a background of working in the wholesale animal or wet market [1]. Other stories, such as the one that claimed that a government agency (e.g., China) manufactured the virus, are still circulating. There have been suggestions that the virus may have come through biological-warfare laboratories. Hoaxes also circulate on how COVID-19 entered humans in the first place and through which animal(s). Although its origin is still unclear, most studies indicate that COVID-19 originated from a viral strain found in bats [2]. From Wuhan, then across China, COVID-19 has now spread to every continent, including Antarctica, and to almost

* Supported by Texas A&M, San Antonio.

every country in the world. Some people try to differentiate between "first recorded outbreak" and "the origin." Many believe that although the first recorded outbreaks were in Wuhan, the origin of the virus is yet unknown.
- The statistics: Different countries report their COVID-19 statistics in terms of reported cases—new, recovered, and deaths. Nonetheless, the accuracy of those numbers is questioned. Different factors can be behind why reported numbers are not accurate. For example, to avoid adverse public response, some governments may not report actual infected cases. This was the case in 1918, when the world was at war and countries did not want to report casualties for fear of the information being used against them. In Europe, Spain was the only country to report deaths. Ironically, Spain was identified incorrectly as the influenza source—hence the name "Spanish Flu" [3]. Similarly, there may be problems with the availability and accuracy of COVID-19 test facilities. Additionally, several reports indicate accuracy problems with test results and that in a certain period of the disease or for certain humans, symptoms may not be visible.
- Possible treatments: At the time this paper was written, there was no confirmed treatment for COVID-19, although, since the beginning of the pandemic, people worldwide started trying different types of treatments, including chemicals in response to the misinformation they received or learned through television, the Internet, or online social networks (OSNs). Some opportunists took advantage of human fears and eagerness to find a treatment and started making their own and selling them through online stores. Other widespread rumors about possible treatments included food ingredients, vitamins [4], and even chemicals (e.g., internal injection of disinfectants). For example, Snopes listed rumors about COVID-19 treatments that included claims related to some natural foods such as bananas, garlic, and lemons.

Lack of credible information on such worldwide issues can create large-scale false information. Recent years have witnessed the large-scale spreading of fake news and rumors campaigned by government agencies, such as the Russian Troll Factory in St. Petersburg. In particular, between the United States and China, misinformation about the origin of COVID-19 added to already-strained relations. Rumors spread across the two countries, where each was trying to connect the roots of COVID-19 to the other country. The following questions guided our research:

- With the variation in how researchers define what is fake and what is true in COVID-19 news and misinformation in general, how can we use transfer learning from one misinformation dataset to another or from one classification model to another?
- While reporting false claims for fact-checking websites can be straightforward, how can we report true claims? How can we differentiate between (1) irrelevant claims—those that can be true or false but are not related to COVID-19 and (2) true claims—those that are both related to COVID-19 and true? This can create two types of datasets:

- Claim/no claim, where the subject will be either a true COVID-19 claim or not. A false claim can be interpreted in two ways: It could be misinformation or irrelevant information to COVID-19.
- False vs. correct COVID-19 claim: false relevant claim versus true relevant claim).

2 Experiments and Analysis

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

– COVIEWED 2020, (https://www.coviewed.org/) [7].
– CoAID (COVID-19 heAlthcare mIsinformation Dataset) [6].
– FakeCOVID [8].

We noticed that those three datasets have different interpretations of what is false and what is true. While we acknowledge that classifiers' accuracy would be impacted with such integration, we nonetheless believe that this combination of datasets will achieve two main goals: Produce models that are capable of working in unknown territories (e.g., using new datasets or claims). Reduce possible bias in proposed models. Classification models that are generated based on specific datasets may be biased or overfitted. In other words, they may work well in the evaluated datasets but poor in any other dataset.

2.1 Combination of COVID-19 misinformation datasets

We aggregated data about COVID-19 misinformation from different sources for several reasons: Different datasets have different conventions and approaches to defining what is true or false in COVID-19 misinformation and how they label each article or claim. One typical problem with using a single dataset is 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 data models work well in a dataset or a subset of a dataset and work 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 datasets in both training and testing. The combined dataset has a total of 20,563 claims, 7,905 of which are false claims and 12,658 of which are true claims. Preliminary text analysis for the false versus true text showed some differences. The first feature we evaluated was the word count.

2.2 Text 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 1 shows an assessment of using CV for several classifiers.

For our four evaluated performance metrics, Precision, Recall, F1-Score, and Accuracy, with the exception of KNN, most classifiers have values between 70% to 80% in all those metrics. Again with the exception of KNN, all classifiers showed similar values in all metrics between false and true claims. Using TF/IDF, Figure 2 shows similar results to those using count vectors.

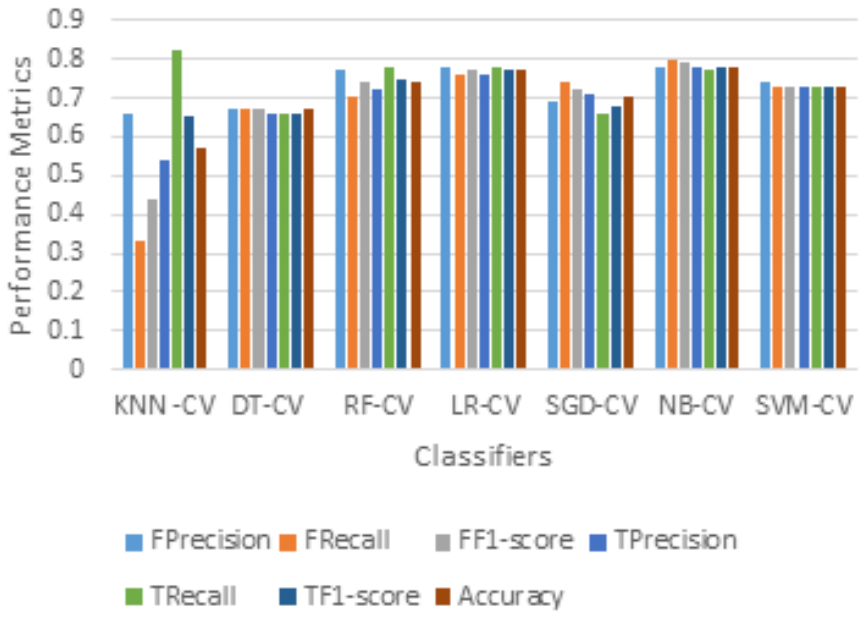

Fig. 1. Classifiers' performance metrics: count vector, 100 terms, no embedding.

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 15000 claims while ensuring that the same number of false and true claims were 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 one model, CV text-feature extraction. Table 1 summarizes results from evaluating the number of terms on classifiers' performance metrics.

The table summary shows the following: Among 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. All other three evaluated classi-

Fig. 2. Classifiers' performance metrics: TF/IDF, 100 terms, no embedding

Table 1. 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	50000	0.88	0.79	0.83	0.81	0.90	0.85	0.84
SGD1	50000	0.89	0.79	0.84	0.81	0.90	0.86	0.85
SVC1	50000	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	50000	0.88	0.80	0.84	0.82	0.90	0.86	0.85
SGD2	50000	0.89	0.81	0.85	0.83	0.90	0.86	0.86
SVC3	50000	0.87	0.82	0.84	0.83	0.88	0.85	0.85

fiers 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.

2.3 Learning with word-embedding models

Word/sentence embedding is a method to obtain a context-dependent vectorized representation for every word/sentence in the text corpus. This representation allows comparison of words in embedding space: Words closely spaced together have a similar meaning and/or connotation, while 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 corpora such as Wikinews, news articles, Google News, BERT, and so on. Recent state-of-the-art word-embedding models such as BERT have proven to be very good at 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 [10]. 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 [11]. In order 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 claims, both were trained with the BERT embedding model. The trained outputs were used as input for all classifiers. Figure 3 shows a summary of the accuracy metric for all classifiers. With the exception of Decision Tree models, all other classification models showed improvement in all performance metrics when using embedding models. Unlike in previous experiments, all classifiers showed no sensitivity to an increase in the model's number of terms.

2.4 Most-informative features

Classification and prediction models are built 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, ngrams, and the like—as features. The process starts with all text. Different preprocessing steps such as stop-words 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 size limitation, we show results from one experiment, TFIDF, most informative features, Table 2. Table 2

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.

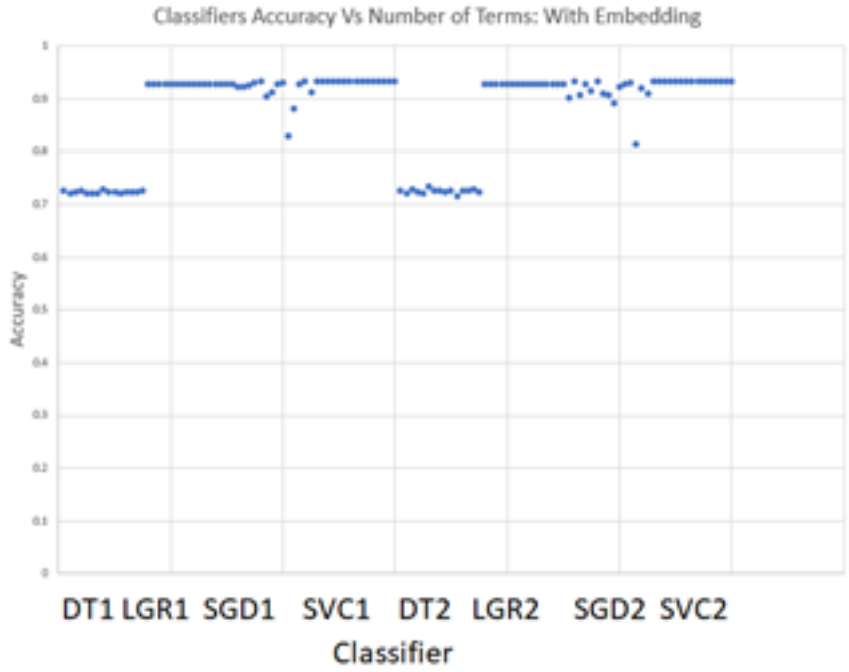

Fig. 3. Classifiers' accuracy versus the number of terms, with embedding.

Table 2. Most informative features: TFIDF.

FALSE	FALSE	TRUE	TRUE
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

A large majority 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 [12], informative terms in the true-claims section include terms that typically exist in news articles, whereas informative terms in the false-claims section include highly specialized terms, indicating that they refer to specific conspiracy theories. Table 3 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 of making a prediction of an instance target label.

Table 3. Top terms using LR and Chi Square.

Term	Chi Square	Term	Chi Square
video	100.15	alongside	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

2.5 Word-embedding models' comparison

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 that we used were W2V, Glove, Google, Paragram, Wiki, and BERT. Overall, SGD and Logistic Regression classifiers scored the best accuracy of values—between 86-87% in most embedding models—whereas the MLP Classifier scored the lowest in most experiments (Figure 4).

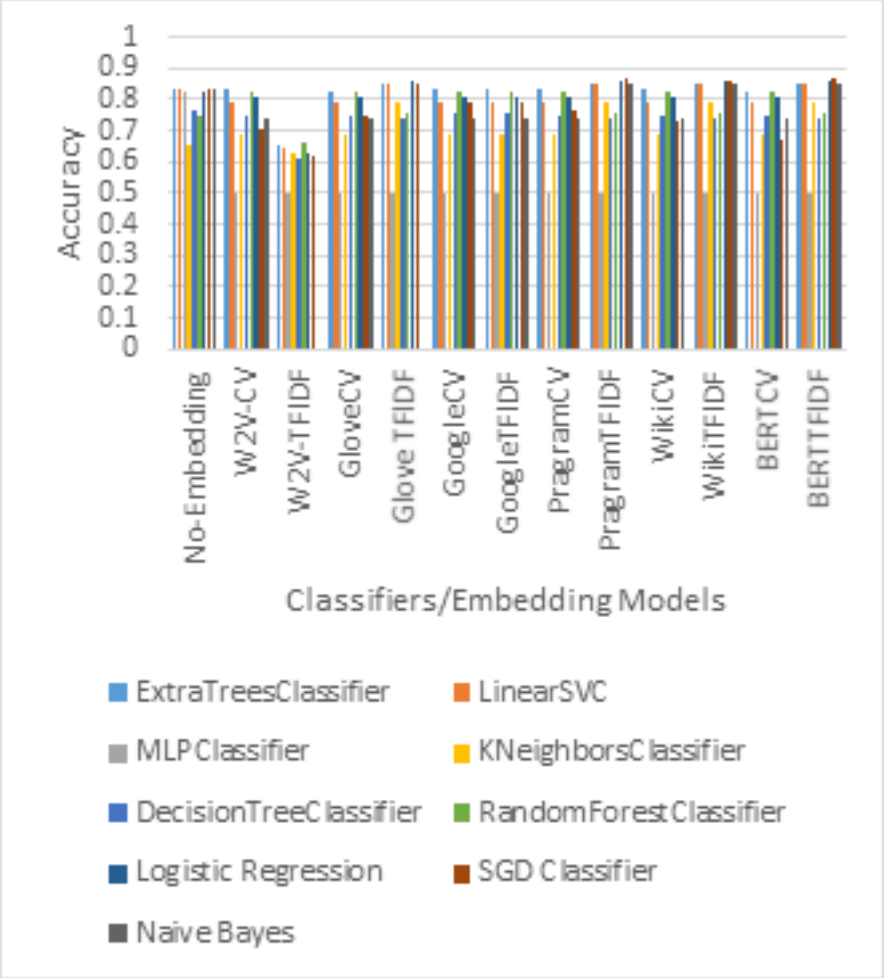

Fig. 4. Classifiers versus word-embedding models.

3 CONCLUSION

With the existence of large sources of information through the Internet, and with the ability of every user to broadcast, how can we verify the correctness or validity of the information? No doubt that this is an open-ended question and no simple solutions are available to fact check such a tremendous amount of information. 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 aspects. As a result, analytic models that are produced based on those datasets can be biased and may not work well in different datasets. We combined several misinformation-related datasets that discuss different aspects of misinformation related to COVID-19. We focused our analysis on 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 input 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.

References

1. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497-506.
2. T. Li, C. Wei, W. Li, F. Hongwei, J. Shi, Beijing Union Medical College Hospital on "pneumonia of novel coronavirus infection" diagnosis and treatment proposal, 2020. (V2. 0). Med J Peking Union Med Coll Hosp.
3. A. Trilla, C. Daer, The 1918 Spanish flu in Spain. Clin Infect Dis. 2020. 47(5), 668–673.
4. R. Gallotti, F. Valle, N. Castaldo, P. Sacco, M. De Domenico, Assessing the risks of infodemics in response to COVID-19 epidemics. 2020. arXiv preprint arXiv:2004.03997.
5. I. Alsmadi, M. O'Brien, How many bots in Russian troll tweets?. Inform Proc Manage, 2020. 57(6), 102303.
6. L. Cui, D. Lee, CoAID: COVID-19 Healthcare Misinformation Dataset. 2020. arXiv preprint arXiv:2006.00885.
7. COVIEWED. 2020. Kaggle.com. Retrieved November 29, 2020, from https://www.kaggle.com/trtmio/project-coviewed-subreddit-coronavirus-news-corpus
8. G. Shahi, D. Nandini, FakeCOVID—a multilingual cross-domain fact check news dataset for COVID-19. 2020. arXiv preprint arXiv:2006.11343.
9. IDeaS, Center for Informed Democracy & Social - cybersecurity (IDeaS) - Carnegie Mellon University. Retrieved November 2020 from https://www.cmu.edu/ideas-social-cybersecurity/research/coronavirus.html.
10. N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 2019. arXiv preprint arXiv:1908.10084.

11. P. Huilgol, Sentence Embedding Techniques One Should Know— With Python Codes. Analytics Vidhya. 2020. https://www.analyticsvidhya.com/blog/2020/08/top-4-sentence-embedding-techniques-using-python/.
12. J. Fairbanks, N. Fitch, N. Knauf, E. Briscoe, Credibility assessment in the news: do we need to read. 2018. In Proc. of the MIS2 Workshop held in conjunction with 11th Int'l Conf. on Web Search and Data Mining (pp. 799–800).