Classifying COVID-19 misinformation with machine learning techniques

Matthew H Goodman
Computer Science
Drexel University
Philadelphia, United States of America
mhg46@drexel.edu

Jeffrey Braga
Computer Science
Drexel University
Philadelphia, United States of America
jjb427@drexel.edu

Doris Chia-ching Lin
Computer Science
Drexel University
Philadelphia, United States of America
dl963@drexel.edu

***Abstract*—Within the context of the COVID-19 pandemic, a novel wave of false information continues to spread regarding new vocabulary, reports, medical guidance, and statistics that are highly specific to the current situation. In this paper, we investigate the use of false news classification through algorithmic application of Logistic Regression and Random Forests. Finally, with an evaluation against purely COVID-19 related articles, we demonstrate that the problem of false information classification, even information which includes unique circumstances and terminology, is an achievable task by achieving an F-Measure of 92% and 72% respective to each algorithm.**

***Index Terms*—classification, logistic regression, natural language processing, fake news**

I. Background

Though the spread of fake news has been the subject of previous research in Artificial Intelligence, news surrounding the pandemic is unique in that articles related to the pandemic are characterized by terminology specific to COVID-19 as well as new rules and routines, social distancing guidelines, public health officials, treatments, and more which result directly from current circumstances.

Information presented to individuals in the current environment will directly impact the choices they make regarding their own health. Furthermore, this is a situation where we must acknowledge that both societal choices and our own personal choices affect the well-being and possible infection of others [15]. As such, the ability to distinguish between the true and false information about health, safety, and other COVID-19 related news would be an invaluable tool. Within the short time-frame since the global crisis began, there have been conflicting recommendations on social distancing, different remarks on the origin of the virus, dubious claims of miracle cures and unorthodox treatments, and even those who claim the virus itself is no more than a hoax [16]. Accordingly, this new wave of false information can be said to impact both public and personal health in a direct sense.

We sought the use of Natural Language Processing (NLP) to build a classification system for accurately separating true articles from those which are false. As NLP is geared towards making sense of human language, the problem of "making sense" becomes much more interesting with the addition of new vocabulary like the terms related to the pandemic. Of the variety of algorithms available for classification, our focus is on the application of logistic regression and random forests. The system is trained on news articles occurring before the timeframe of the pandemic, primarily relating to news of a political nature and the 2016 United States Presidential election. Afterwards, it is then evaluated in its ability to correctly classify articles relating to COVID-19 pandemic, which were collected by our team as the news cycle relating to the crisis unfolds in real time.

II. Related Work

Classification can be understood as the task of predicting the correct category or value of a given data point. This has been studied in both supervised and unsupervised machine learning approaches, and an assortment of algorithms have been proposed to best accomplish this kind of prediction. In closer relation to our topic, Deep learning literature contains varying studies on classifying "fake news" [1]–[3]. Most of these studies have tested multiple algorithms to find which best fits a specific problem, and the accuracy measurements for these studies range from 50% to 92%. Additionally, to identify instances fake news, some have recommended to denote between reliable and unreliable news with source, content, and subject [5]. Perhaps as the result of a desire to study the classification of truth and falsehood in a more broad sense, the current work in the field primarily focuses on classifying "fake news" in general. Our study hopes to expand the literature by focusing on classifying "fake news" surrounding a specific news topic, COVID-19.

Currently, the COVID-19 pandemic has come to dominate the media, both domestically and abroad. The spread of fake news offers unique challenges and dangers to the public. It appears certain political groups and state agents may wish to exploit the situation to push their own image [4]. In addition, if the information is coming from sources (especially governmental sources) that people consider reliable and trustworthy, spread of such unreliable information creates greater danger to the public. With the unknown and uncontrollable quality of the situation during the pandemic, people seek safety, security and order to compensate for lack of control and our sense of isolation as a result of quarantining and social distancing measures. This may also make some individuals more likely to accept information without scrutiny.

This specific topic for false news presents its own dangers with rates of inaccurate classification. This becomes evident in the same sense in which a misdiagnosis for a medical situation is more dangerous than a system incorrectly classifying a photo of a dog into a corresponding category of images. Therefore, a high rate of false positives and negatives resulting from an evaluation of our model would not only fail to accomplish our intent but also unintentionally present a means of classification which could dangerously increase the spread of misinformation we are seeking to debase.

III. Methodology

All programming was done in Python and made use of the Natural Language Toolkit [9] and Scikit-Learn [10] packages. To classify fake news, our team used two datasets. The first dataset was used to train our models. The second dataset was used strictly for testing purposes.

A. Train Dataset

We obtained the training dataset from Kaggle [6]. The dataset contains 21,417 articles labeled as "true" news and 23,502 articles labeled as "fake" news. The article originated from accredited news sources and contained topics such as politics, world news, and local news. To add familiarization with the COVID-19 topic, our team gathered an additional 81 true COVID-19 articles and 9 fake COVID-19 articles to add to the dataset. After adding the additional articles, the sample size was 21,498 true articles and 23,511 fake articles.

All articles were formatted in comma-separated values (CSV), such that each article contained four columns of data: 1) article title; 2) article body text; 3) article subject; and 4) article publishing date. Article text only consisted of ASCII characters and were limited to the English language.

The training dataset contained 21 articles with missing values. Our team decided to remove those articles from model training. The final sample size for the training dataset was 21,498 true articles and 23,490 fake articles.

B. Test Dataset

Our research team gathered an additional 100 COVID-19 articles from various sources. The dataset consisted of 91 true articles and 9 fake articles. To determine if an article presented fake news, the article was cross-referenced with fact checking resources, such as Agence France-Presse [14], one of the oldest international news agencies which documents instances of fake and misleading news and archives articles so that they may still be viewed even if the website is taken down. A vast majority of false news documented on Agence France-Presse regarding COVID-19 were social media posts rather than news stories, but for our purposes we have only used those published as articles in keeping with our training dataset. We note that the test dataset did not contain missing values.

C. Data Preparation

To pre-process the data for both training and testing, our team first merged the article title, text body, and subject into a single data field. Once merged, all characters were converted to lowercase letters.

Following the advice of Srivasatava [13], we removed stop words and stemmed each text sample. Stop words are words that serve primarily for syntactic function, such that they do not impact the subject matter [7]. Some examples include, "a", "how", and "or". The other method, stemming, involves removing the root word from a word [13]. For example, the word "fishing" would be analyzed as "fish".

D. Training and Testing the Model

To train the classifier, our team used a 80% train and 20% test split on the training dataset. We decided to assess the performance of the classifier on the training dataset to ensure the model would be capable of successfully classifying a new and foreign dataset.

Two algorithms were implemented to build the models. First, the data was trained and assessed on a logistic regression algorithm. Logistic regression is a model that determines which of a set of classes is more probable given a set of data features [12], and the probability model for it is defined as:

$$P(Y_n = 1|x_n) = \sigma(w * x_n)$$

The second algorithm was random forests. Random forests use multiple random trees for classification [8]. In the set of trees, each tree is given an equal weight vote, and the label with the most votes is selected as the prediction.

Once the classifiers were developed, the COVID test dataset was run through each model, and the performances were assessed.

IV. Results

We assessed the classifiers with three performance metrics: 1) recall; 2) precision; and 3) f-measure. For the 80/20 split on the training dataset, both models performed similarly. The metrics for the training dataset are shown in Table I, and the metrics for the COVID test dataset are shown in table II.

Furthermore, our team plotted confusion matrices for additional insights into the classifier's performance. These are shown in Fig. 1 for logistic regression and Fig. 2 for random forests.

Fig. 1. Confusion Matrix for Logistic Regression

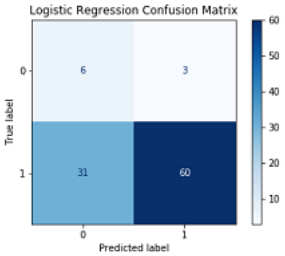

Fig. 2. Confusion Matrix for Random Forests

TABLE I
PERFORMANCE METRICS ON THE TRAINING DATASET

Algorithm	Precision	Recall	F-Measure
Logistic Regression	0.9907	0.9928	0.9918
Random Forests	0.9967	0.9949	0.9958

V. CONCLUSION

Our implementation further demonstrates that news classification is a worthwhile task for the field of Artificial Intelligence, even regarding new and novel subject areas like COVID-19 that continue to arise in our world. Within our proposal, we mentioned that a classifier with an accuracy rate of 50% would be no better than an uneducated guess or coin flip mentality of classifying articles. We have far exceeded that level of accuracy with our model. That being said, our model is by no means perfect, and there is still room for improvement and greater accuracy.

It is also worthy of attention that the pandemic is still occurring at the time of this paper's completion. Though the number of true articles referencing COVID-19 will undoubtedly increase as time passes, we cannot predict the quantity or content of the sorts of false information that will spread in the future. Because we are evaluating this in real time rather than as a past event, it would be advantageous to continually evaluate this model against new falsehoods to see if misinformation is becoming harder to track or seemingly more and more "real" to readers. Though it was a noted difficulty for our group to track down archived false articles, this difficulty is a positive for the situation as a whole. However, there is much more nuanced and grey area to be explored in this field, where information conveyed is often more of a misleading or misinterpreting nature in a subtle way than flat out false.

One particular area of possible future research we want to strongly highlight is the role of social media in spreading false news. In the process of searching for false articles related to COVID-19 on Agence France-Presse, an overwhelming amount of documented falsehoods were spread via Facebook and Twitter in the format of posts, tweets and infographics. In fact, many have suggested that social media is the preferred method of misinformation spread for those with malicious intent.

At the outset of our project proposal, we did not consider a dataset with a foundation centered on this problem. With this means of spreading false information growing in the modern world, it would be an immense leap forward for any classifier of fake news to have data from social media outlets for the purpose of training a model. Pla and his team [11], have proposed a method of automating the creation of such a dataset as opposed to relying on human annotation, which can create other unforeseen problems. The most worthwhile endeavor in furthering our research would likely be this sort of additional data.

TABLE II
PERFORMANCE METRICS ON THE COVID-19 TEST DATASET

Algorithm	Precision	Recall	F-Measure
Logistic Regression	0.9529	0.8901	0.9205
Random Forests	0.0.9524	0.0.6593	0.7792

ACKNOWLEDGMENT

REFERENCES

[1] Bajaj, S. (2017). The Pope Has a New Baby! Fake News Detection Using Deep Learning. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., ... Scala, A. (2020). The covid-19 social media infodemic. arXiv preprint arXiv:2003.05004.

[2] Gilda, S. (2017, December). Evaluating machine learning algorithms for fake news detection. *In 2017 IEEE 15th Student Conference on Research and Development (SCOReD)* (pp. 110-115). IEEE.

[3] Liu, Y., Wu, Y. F. B. (2018, April). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *In Thirty-Second AAAI Conference on Artificial Intelligence*.

[4] Wanat, Z., Cerulus, L., Scott, M. (2020, March). EU warns of "pro-Kremlin" disinfo on coronavirus pandemic. *In Politico*.

[5] De Witte, Melissa (2020, March). Why fake news about coronavirus is appealing (and how to avoid it). *In Phys.org*.

[6] Bisaillon, C. (2020, March). Fake and real news dataset, Version 1. Retrieved May 25, 2020 from https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset/metadata

[7] El-Khair, I. A. (2017). Effects of stop words elimination for Arabic information retrieval: a comparative study. arXiv preprint arXiv:1702.01925.

[8] Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, 1-13.

[9] Loper, E., Bird, E. (2002). NLTK: The Natural Language Toolkit. In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pages 62–69. *Association for Computational Linguistics*.

[10] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.

[11] Pla Karidi D., Nakos H., Stavrakas Y. (2019) Automatic Ground Truth Dataset Creation for Fake News Detection in Social Media. Intelligent Data Engineering and Automated Learning – IDEAL 2019. IDEAL 2019. *Lecture Notes in Computer Science*, vol 11871. Springer, Cham

[12] Schein, A. I., Ungar, L. H. (2007). Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3), 235-265.

[13] Srivasatava, S. K., Kumari, R., Singh, S. K. (2017, May). An ensemble based nlp feature assessment in binary classification. *In 2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 345-349). IEEE.

[14] Plateforme de l'info. (2020, March 12). Retrieved from https://www.afp.com/

[15] Everyone Included: Social Impact of COVID-19 — DISD. www.un.org/development/desa/dspd/everyone-included-covid-19.html.

[16] Finnegan, Conor. "False Claims about Sources of Coronavirus Cause Spat between the US, China." ABC News, ABC News Network, 12 Mar. 2020, abcnews.go.com/Politics/false-claims-sources-coronavirus-spat-us-china/story?id=69580990