STEM Fellowship Big Data Challenge 2021

Covid-19 Fake News Detection

Yuyan Chen[1], Zijun Zhao[1], Dayou Ren[2], and Xichen He[3]

[1]McGill University
[2]University of Texas at Austin
[3]University of California, Santa Barbara

May 31, 2021

Abstract

Fake news detection is crucial for addressing misinformation on social media. This project aims to propose machine learning models for fake news detection based on text mining. Three models, naive Bayes, logistic regression, and random forest, are trained based on term-frequency-inverse document frequency (TF-IDF). The effectiveness of text mining-based models is justified with model accuracy. Possible improvements for future studies are suggested based on the discussion of limitations of text-mining based models.

Keywords
Covid-19, fake news, NLP

1 Introduction

In the course of fighting against COVID-19, misinformation and disinformation are significant hindrances to enforcing public health measures. News reports are an influential source of misinformation, and tools that assist in detecting the authenticity of news can help control the spread of misinformation. Given the tremendous amount of news and misinformation, it is impossible for human beings to easily detect fake news at first glance, while machine learning, data-driven and efficient, can provide a solution for this task. Using machine learning for text classification has been studied for decades, and many methods and statistics for natural language have been developed. This project analyzes the text of fake news and real news and mainly uses term frequency-inverse document frequency (TF-IDF) as a numerical statistic to quantify the analysis. Models, including Naive Bayes, Random Forest, and Logistic Regression, are trained based on TF-IDF. The accuracy of the models is explained with word clouds and TF-IDF distribution, and the effectiveness of this project is justified with the high accuracy of the models.

2 Materials & Methods

Two datasets, consisting of 10,785 pieces of fake news and 2535 real news, were used for this project: COVID-19 Fake News Dataset[1] and COVID Fake News Dataset[2].
Datasets were preprocessed with Natural Language Toolkit [3] including removing stopwords and punctuation, coverting all letters to lowercase, and tokenization. Word frequencies were calculated using Term Frequency - Inverse Document Frequency (TF-IDF) and 10,000 features were extracted using scikit-learn TfidfVectorizer with unigrams and bigrams[4].
Three models were trained with the 10,000 features: Naive Bayes, Random Forest, and Logistic Regression. The Naive Bayes model is implemented as a mutilnomial Naive Bayes model as it is a classic model used for text classification. The minimum number of samples required to split an internal node is set as 2 for the Random Forest model and nodes are spilt until all leaves are pure or contain less than 2 samples. This repository contains the preprocessed dataset used for this project as well as all the source code.

3 Results

Overall, the Random Forest classifier gives the highest accuracy (93%) and Naive Bayes and Logistic Regression both give 92%. 94% fake news and 81% real news are correctly classified by Naive Bayes classifier. 95% fake news and 83% real news are correctly classified by Random Forest classifier. 93% fake news and 83% real news are correctly classified by Logistic Regression classifier.

The classification report and confusion matrix (in percentage) for three models are as follows:

Table 1: Classification Report for Naive Bayes

	precision	recall	f1-score	support
fake	0.94	0.96	0.95	2171
real	0.81	0.71	0.76	493
accuracy			0.92	2664
macro avg	0.87	0.84	0.85	2664
weighted avg	0.91	0.92	0.91	2664

Figure 1: Confusion Matrix for Naive Bayes

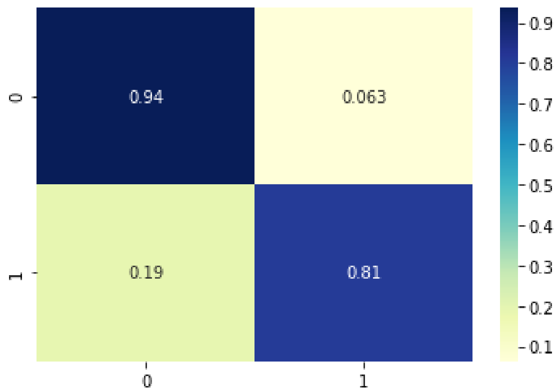


Table 2: Classification report for Random Forest

	precision	recall	f1-score	support
fake	0.95	0.96	0.96	2171
real	0.83	0.76	0.79	493
accuracy			0.93	2664
macro avg	0.89	0.86	0.88	2664
weighted avg	0.93	0.93	0.93	2664

Figure 2: Confusion Matrix for Random Forest

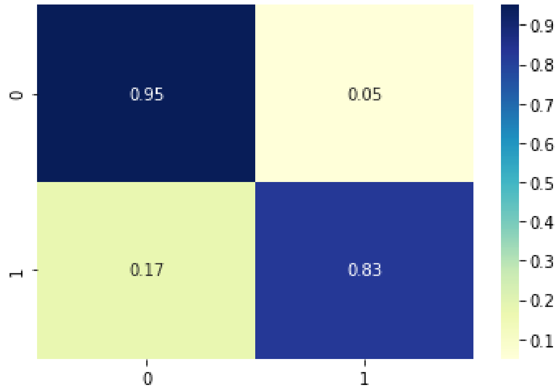
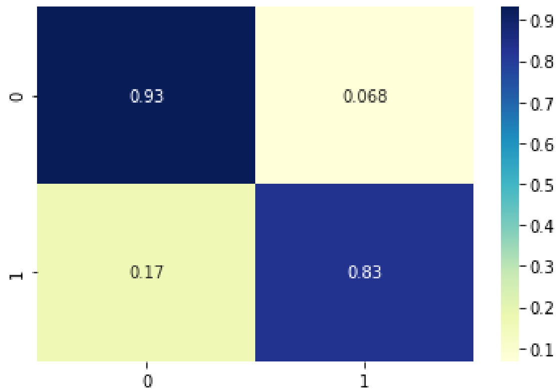


Table 3: Classification report for Logistic Regression

	precision	recall	f1-score	support
fake	0.93	0.97	0.95	2171
real	0.83	0.69	0.75	493
accuracy			0.92	2664
macro avg	0.88	0.83	0.85	2664
weighted avg	0.91	0.92	0.91	2664

Figure 3: Confusion Matrix for Logistic Regression

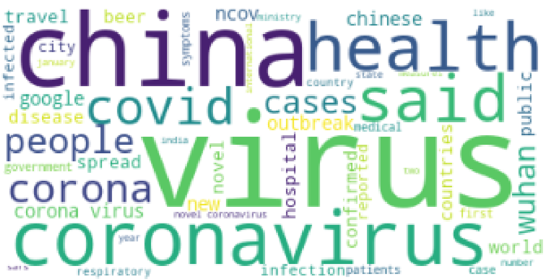


4 Discussion

All three models are trained with 10,000 features extracted with TF-IDF, and all the models give high accuracy (over 90%), which suggests an obvious difference in wording between fake news and real news and fake news detection based on text mining only is practical. For example, the word clouds below clearly show the difference in word choices between fake news and real news.

Figure 4: Word Cloud of Fake News

Figure 5: Word Cloud of Real News



Summing the TF-IDF scores of each word in fake news and real news respectively shows the specificity of each word to fake news and real news. The top 10 words in fake news (set1) are entirely different from the top 10 words in real news (set2). The top 10 words or phrases (sorted with TF-IDF sums) in real news (set2) rank relatively low in fake news and vice versa. Figure 7 below clearly indicates that the TF-IDF sum distribution of set2 calculated from fake news is entirely different from that calculated from real news.

Table 4: Rank of Words

set1	rank real	set2	rank fake
contract	3106	virus	3037
coronaviruses zoonotic	9590	china	4105
peer	8177	coronavirus	4192
villages	7547	health	3062
china may	6343	said	3508
virus found	3705	covid	2432
showing symptoms	1823	corona	94
newspapers	4286	people	1144
lost	1991	cases	8285
pakistani	1782	wuhan	8108

Figure 6: Top10 Words in Fake News

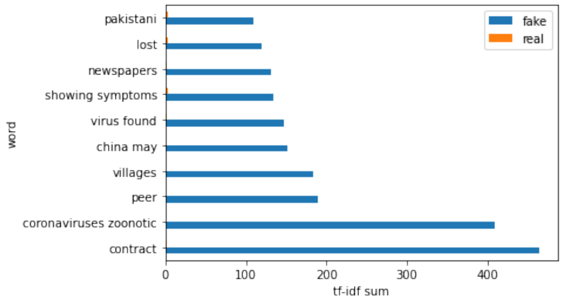
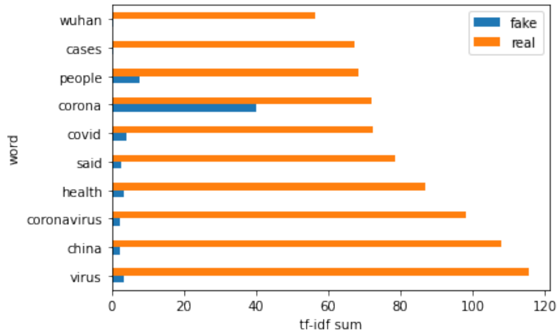


Figure 7: Top10 Words in Real News



It is expected that all three models are more sensitive to detect fake news since the training set is imbalanced and the number of fake news is more than four times the real news. Sensitivity to fake news corresponds to the purpose of this project, as fake news that would be classified as real by the models can cause more harm than its counterpart. The huge difference in rankings explains the high accuracy of the three models and corresponds to the performance of the multinomial Naive Bayes model, which is a classic model for text classification.

The percentage of correctly classified news with over 90% confidence differs greatly among the three models. The Naive Bayes classifier correctly classifies 82.42% fake news and 61.36% real new with over 90% confidence, while the Random forest classifier and logistic regression classifier only correctly classifies 9.87% and 18.88% with over 90% confidence respectively.

However, these text-mining based models have the following limitations. While choice of words has a strong correlation with the the authenticity of a piece of news as suggested, it

Table 5: Correctly Classified News with over 90% Confidence (Percentage)

Model	fake	real
Naive Bayes	0.8242	0.6136
Random Forest	0.8411	0.0987
Logistic Regression	0.8535	0.1888

is not a solid evidence to determine whether a piece of news is fake or not. It is hard for human to observe a pattern in fake news as people do not detect fake news only based on the words shown in Figure 4 and Figure 5. Also, journalists can also write fake news that can trick these models by following the structure and word choice of real news and replace the facts with false information. Besides, the performance of these models on a piece of news highly depends the similarity of this piece to the news in the dataset. Therefore, these models have to be trained frequently with up-to-date news to maintain high accuracy. A common method for human to detect fake news is to compare the piece with credible sources and then determine if the news contains any misinformation or disinformation. While training models based on text is simpler and also efficient, training models based on checking sources might offer more useful information of users.

Conclusions

Fake news detection is still being researched and developed in many different ways. Analyzing text with term frequency-inverse document frequency is one of the simplest methods to detect fake news. Its high accuracy is justified with the dramatic difference in TF-IDF distribution and word ranking. Together with source checking, these models can be used to check the authenticity of news and help users to roughly filter out the fake news. Future studies can incorporate models based on source-checking into these text-mining based models

Acknowledgements

We would like to express our gratitude to the organizers and mentors of 2021 Big Data Challenge.

References

[1] Koirala, Abhishek (2021), “COVID-19 Fake News Dataset”, Mendeley Data, V1, doi: 10.17632/zwfdmp5syg.1

[2] Sumit Banik. (2020). COVID Fake News Dataset [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4282522

[3] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.