Disinformation/ Fake News detection

Social Media Information Classification During The Year of COVID-19 and Election

Shangbin Tang
Department of Geology and Environmental Science
University of Pittsburgh
Pittsburgh Pennsylvania USA
shangbin.tang@pitt.edu

Chengchen Wang
Joseph M. Katz Graduate School of Business
University of Pittsburgh
Pittsburgh Pennsylvania USA
chw183@pitt.edu

Taylor Herb
School of Computing and Information
University of Pittsburgh
Pittsburgh Pennsylvania USA
tah77@pitt.edu

ABSTRACT

The pervasiveness of social media has given rise to a new wave of information explosion. Users and social media platforms both need mechanisms for generally classifying the authenticity of the information in order to indicate possible misinformation.

Given the short-term topic trending and time-sensitivity of social media messages, this project used social media data since January 2020 to evaluate the algorithms' effectiveness for short-term message classification. In addition to text length, sentiment tendency, and DTM, this project introduced credit scores based on the author's historical posting statistics to evaluate and classify the authenticity of messages. The project also compared the efficiency and accuracy of several classification algorithms.

The results show that the use of credit scores can improve the accuracy of classification; the algorithm of Random Forest has the highest accuracy among the algorithms applied in this project. This project demonstrates that user's historical authenticity statistics as credits can effectively improve the accuracy of social media information authenticity classification and have the potential to be utilized in social media and applications' backend systems.

KEYWORDS

Data Mining, Text Mining, Text Classification, Social Media, Fake News

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 INTRODUCTION

Following the appearance of television and the Internet, the widespread use of social software has led to a new information explosion. Social media enables people to make the most of their freedom of speech, and individual voices can have a larger audience than ever before. However, unlike traditional media, new media lacks a complete and well-placed fact-checking and monitoring mechanism. Consequently, new media has also become a platform for the co-existence of true and false information. Meanwhile, the popularity of cell phones and computers provide social media with a wide range of users with different cognitive abilities, experiences, and the ability to distinguish between true and false information. Under these circumstances, misinformation sent by social media influencers can mislead a larger number of people.

The year 2020 is an unusual year with many events of significant and widespread social impact, such as the COVID-19 pandemic, the BLM movement, and the U.S. presidential election. People are expressing and spreading their agendas on social media on these topics, while some people are also posting misinformation about them.

The purpose of this project is to find algorithms and the influencing variables that can better identify and classify true and false information from short messages on social media.

2 DATA AND METHODS

2.1 Data

The data of this project were obtained from a news fact-check website, PolitiFact (https://www.politifact.com/), which provides fact-checking results for some of the most popular posts and quotes on social media. Using python coding with web scrape package: Beautiful Soup, we obtained all of the site's fact-checked and published message logs from January to November 2020. The obtained records include author, time of posting, information about the occasion (platform) of posting, and message content. After data pre-processing and removal of invalid records, a total of 2123 records will be used for the analysis and classification.

The original website used a 6-level system of classification to evaluate the authenticity of messages. In this analysis, we followed this standard but also merged it into a more general 3-level criterion(see Table 1) and performed classification based on both systems to compare the effectiveness of different

classification algorithms by comparing the classification and accuracy evaluation results.

The term-document matrix was built based on processed and cleaned data. The number of words contained in each record was counted as content length, which would participate in the following classification as a variable. The sum of the sentiment scores for each record also took part in the classification.

Compared to other text-based fake information detection methods, this project introduced credit scores, which is the statistics of historical true, half-true and false information published by authors. The exploratory analysis indicated that the number of true and false messages published varies by authors (sources). New messages posted in the future by authors who had published more false information before are more likely to be false. Also, using statistics for the categories of messages posted by the authors instead of the authors themselves avoids bias. Therefore, the number of false and true messages posted by the author of each message is an important variable in the classification analysis. In addition, since the data were from various occasions, social media, and platforms, the historical true/false statistics of each occasion would also be included in the following classification.

2.2 Methods

2.2.1 Data Cleaning. To get the data for classification, we first checked the data's validity and usability by viewing the dimensions and summary of the dataset. Since we want to train two types of categories with and without author and occasion's credit score, we split the rearrange the original dataset into four datasets. The structure of all variables is character, and we get the clean dataset by dropping missing values and converting the necessary predictor's category to factor. Based on the document-term-matrix, the indicator of the content length and sentiment score of each document is calculated. The high-frequency (frequency is higher than three) document-term-matrix is joined in the dataset to keep the document context as much as possible and save the running time.

Table 1: Authenticity Level Categories

Original Authenticity Level Categories	Merged Authenticity Level Categories
True	True
Mostly-true	True
Half-true	Half-true
Mostly-false	False
False	False
Pants-on-fire	False

2.2.2 Term-Document Matrix. Since we want to use the content for disinformation detection, we need to extract information from unstructured textual resources and get a document-term matrix. Here, we lowercase all texts, remove stop words, numbers, punctuation, special characters. Then we stem all texts into its roots format and get the clean text corpus. However, because of the limitation of computer speed, we filter out the words that appear less than three times in the whole dataset.

2.2.3 Sentiment analysis. It's believed that false information is more inflammatory and emotional than true information. We want to add a sentiment score to evaluate the emotional inclination of every document for classification. To make a sentiment analysis and quantify the sentiment of every document, the afinn dictionary is used to give a rate for all terms in one document between minus five and plus five. Positive values express active emotion, like joy and happiness, and negative values express bad emotions, such as hate. The

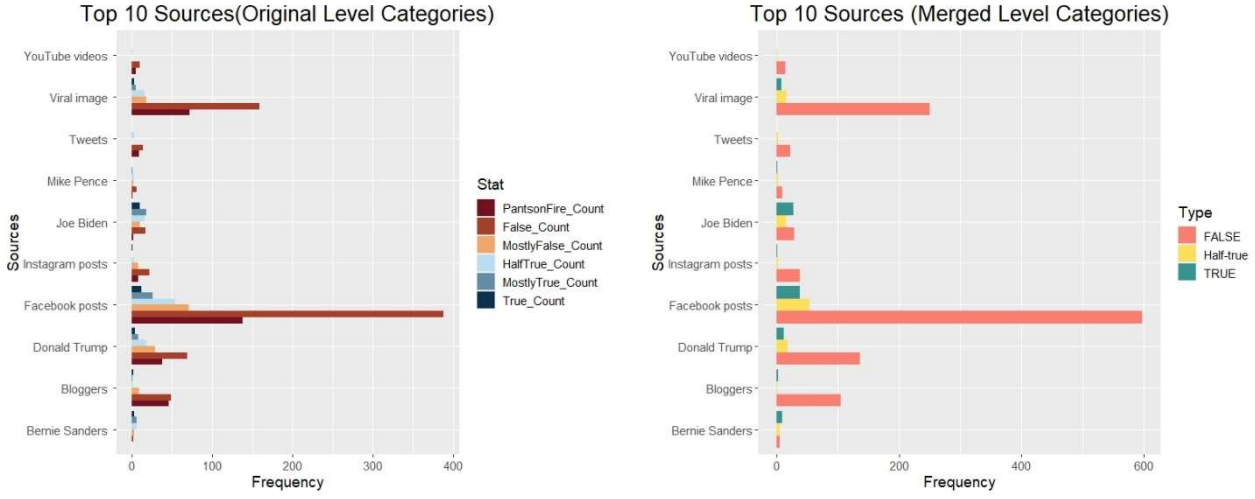

Figure 1: Authors' and Sources' historical authenticity statistic

Table 2: Comparison among classification algorithms

algorithms \ overall accuracy	3-level		6-level	
	with historical score	without historical score	with historical score	without historical score
Random Forest	0.8245	0.7113	0.6302	0.4151
knn	0.7491	0.6774	0.4943	0.3208
svm	0.6962	0.6736	0.3774	0.3774

Table 3: Confusion matrix of classification result with Random Forest algorithm

			Sensitivity	Specificity	Balanced Accuracy
3-level	with historical score	FALSE	0.9894	0.4837	0.7365
		Half-true	0.27273	1	0.63636
		TRUE	0.52874	0.9684	0.74857
	without historical score	FALSE	1	0	0.5
		Half-true	0	1	0.5
		TRUE	0	1	0.5
6-level	with historical score	Pants-on-fire	0.34146	0.93973	0.64
		FALSE	0.9636	0.6258	0.7947
		Mostly-false	0.3333	0.96703	0.65018
		Half-true	0.3939	0.97629	0.68512
		Mostly-true	0.41667	0.97925	0.69796
		TRUE	0.96538	0.96538	0.77756
	without historical score	Pants-on-fire	0	1	0.5
		FALSE	1	0	0.5
		Mostly-false	0	1	0.5
		Half-true	0	1	0.5
		Mostly-true	0	1	0.5
		TRUE	0	1	0.5

higher the rate, the stronger the emotional inclination the word shows. Here, the word that does not show in this dictionary were viewed as having no sentiment inclinations with a rate of zero. By multiplying the frequency of the term by its sentiment score and sum up all sentiment scores of every row, the whole emotional inclination of the document was obtained.

2.2.4 Classification. Each dataset was first split into training and testing sets then cross-validated using 10-fold cross-validation. After resampling, each training set was used to train three classification models. The three classification models used for this project were Random Forest, K-Nearest Neighbor and Support Vector Machine.

3 RESULTS AND EVALUATION

Each classification model was used to predict on the test datasets and the results were evaluated using a confusion matrix (see Table 2). The confusion matrix allowed us to compare the models using the accuracy, sensitivity and specificity metrics. The best performing model was random forest for dataset one.

Dataset one included 3 levels and credit scores. The worst performing model was dataset 4 which included 6 levels and did not include credit scores (see Table 3). Overall, the models trained on datasets including 3 levels performed better than the models trained on datasets including 6 levels. Additionally, the credit score had an impact on the accuracy of the models. The models trained on datasets that included credit scores performed better than models trained on datasets that did not include credit scores. This was true for both 3 level and 6 level datasets.

Unfortunately, it seems there were some issues with the random forest models for the models that did not include the credit score (datasets 3 and 4). The results of the confusion matrix show sensitivity and specificity values of 0 and 1 which is highly unlikely. Therefore, the results of these models should not be trusted.

4 DISCUSSIONS

This project acquired, analyzed, categorized, and evaluated social media data for practical purposes. Credit scores that have not been used in similar projects were introduced. In general, the application of credit scores significantly improved the accuracy of the classification results. Especially for 3-level classification, the results were satisfactory.

However, all algorithms did not classify and identify 6-level data well, especially when credit scores are not included. This may be due to the fact that the differences between 6-level data are smaller compared to 3-level data. Also, since the raw data were manually verified, classified, and leveled, the differences between some categories could be more subtle, or even vague. Algorithms can effectively identify whether information is true or false, but not how true or false it is.

Overall, the classification algorithm is effective in identifying true and false messages and the credit scores contributes a lot. The model cannot identify very accurate or specific how true a message is, but it can give an indication of the overall authenticity of the message. Hence, this model, and especially the credit scoring mechanism, can be used by social media or applications to automatically identify and alert users to possible fake messages and untrustworthy users.

REFERENCES

[1] PolitiFact.com
[2] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data, 8, 3 (Jun 2020), 171-188.
[3] Risdal, M. Getting Real about Fake News. (https://www.kaggle.com/mrisdal/fakenews/notebooks?datasetId=444&language=R)
[4] Chauhan, K. Exploratory Analysis and fake news classification on Buzzfeed News. 2019. (https://www.kaggle.com/kumudchauhan/fake-news-analysis-and-classification)
[5] Dzwolak, M. Fake News Analysis. 2017. (https://www.kaggle.com/michaleczuszek/fake-news-analysis)

Teamwork and Individual Contribution

Shangbin Tang: Participated in discussions on the development of the project title, technical approach, and workflow. Wrote code to capture the data used in this analysis and performed preliminary data organization. Participated in the discovery analysis of the data. Participated in the writing and integration of project documentation.

Chengchen Wang: data cleaning and data pre-processing, sentiment analysis, high frequency word(unigram & bigram) analysis and visualization, powerpoint presentation about data pre-processing and data cleaning and helped write final paper.

Taylor Herb: Ran classification and evaluation analyses on final 4 datasets, wrote classification and evaluation sections of powerpoint presentation and helped write final paper.