Text Classification by using Natural Language Processing

Peiyang Yu[1, *], Victor Y.Cui[2] and Jiaxin Guan[3]

[1]College of Software, Jilin University, Changchun, China
[2]Boston University, Boston, United States
[3]Chengdu No.7 High School, Chengdu, China

*Corresponding author: yupy5517@mails.jlu.edu.cn

Abstract. The spread of Fake news on the Internet misleads people to have false understandings and recognitions of certain events and make ill-advised decisions. This widespread can post a significant threat to modern society politically, economically, and ethnically. In this paper, multiple methods and approaches are mainly discussed for text-data preprocessing, implement different word-vector algorithms. Finally, two datasets are chosen to compare four of the most popular Natural Language Processing (NLP) models to help people distinguish the authenticity of the news.

Keywords: Pre-Processing, Exploratory Data Analysis, Text Classification

1. Introduction

One of the most significant consequences of the rise of the Internet is the faster spread of information worldwide. Although it makes information more accessible to people, it also turns people more vulnerable to false information and fake rumors. Fake news about a company can drastically affect the stock price of such company [1], and fake news posted during a disaster, like during the Hurricane Irma in 2016 [2] and nowadays Covid-19 pandemic, can arise unnecessary tension among people and also potentially put people's lives on risk by obstructing people's correct recognitions of those disasters [3]. Therefore, for the sake of society, it has been a crucial common goal to contaminate the spread of fake news over the Internet.

In this paper, two datasets are chosen from Kaggle Datasets, preprocessed them and evaluate four popular models, respectively Long Short Term Memory Network (LSTM), Multinomial Naïve Bayes (MNB), Gaussian Naïve Bayes (GNB), Random Forest (RF) and Logistic Regression (Log-Reg), on these two datasets and compare the performance of these five models. This comparison is crucial to improve the computer's ability to detect fake news without costing an exhausting amount of human resources, as a better model can usually elevate such capability by a considerable amount.

This paper is organized as follows: In Part 2, the previous related work is reviewed. Then the methodology of our approach is presented in Part 3, followed by the experiment results in Part 4. Finally, Part 5 concludes the full paper.

2. Related Works

Text Classification has been a popular front research topic for quite some time thanks to its extensive and vital applications. Previous studies have shown that despite the requirement of prolonged time and exhausting efforts, supervised classification with manually input features is very useful and important

for detecting false news [4]. There are many supervised models designed for different purposes, but many of the widely used are also well-rounded to suit additional tasks such as the topic of fake news detection in this paper. Long Short-Term Memory (LSTM) was shown to be effective in numerous areas, including acoustic modeling for speech recognition [5], sign language recognition [6], and sentiment analysis [7]. Random Forest has been proven to be a very accurate tool for classification and regression, especially suited for modeling in cheminformatics [8]. Logistic Regression is a traditional statistical model performed well in multiple classifications and regression problems such as Tomography processing [9].

3. Methodology

3.1. Datasets

In this research, two datasets are involved. Both of the two datasets models are downloaded from Kaggle. The dataset 1[10] is a list of online articles collected randomly and classified manually by editors and reporters. The dataset contains 23,481 fake news and 21,417 real news. Figure 1 shows the distribution of classes in dataset 1. '0' is used to represent fake news and '1' to represent Real news. In addition to the balanced dataset 1, an imbalanced dataset 2[11] is also chosen to evaluate the models. It contains 17,014 real job postings and 866 fraud job postings. Figure 2 shows the distribution of classes in dataset 2, with all real postings labeled as 0, and all fraud postings as 1.

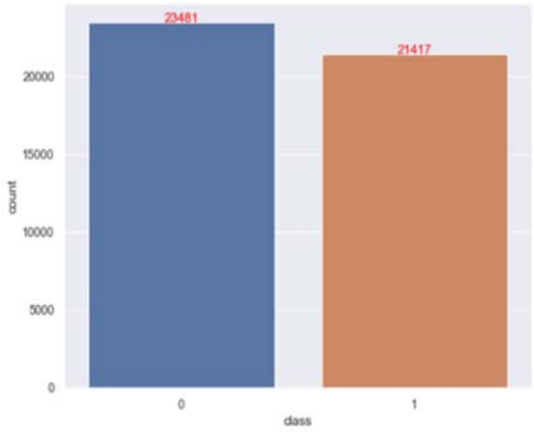

Figure 1. Distribution of Dataset 1.

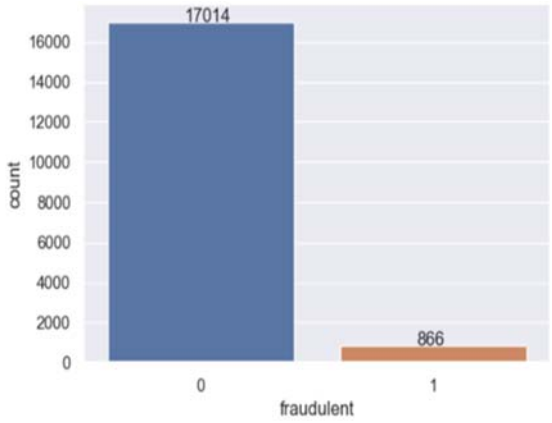

Figure 2. Distribution of Dataset 2.

Balanced datasets like Dataset 1 are usually ideal for training models and will yield optimal results. However, real-world problems often involve extremely unbalanced data like Dataset 2, which is hard to be clean up perfectly. Therefore, our evaluation of the models will be a combination of the five models' performances on both datasets.

3.2. Exploratory Data Analysis

(1) This part mainly consists of the Exploratory Data Analysis (EDA) of dataset 1, then that of dataset 2. The next part will be data preprocessing executed on the two datasets based on the results of the EDA in this part. Firstly, WordCloud is created which can visualize the content of the news/job postings and have a clear view of what appears most frequently in Real/Fake news. Figure 3 and Figure 4 shows the WordCloud of real news and fake news in Dataset 1.

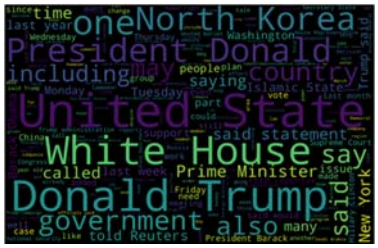

Figure 3. WordCloud of Real News.

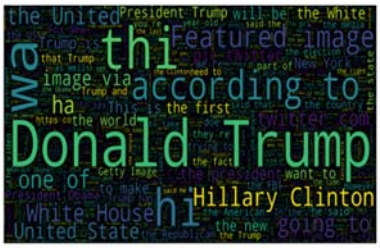

Figure 4. WordCloud of Fake News.

The "title" attribute is also examined. Figure 5 is a boxplot drawn to show and compare the lengths in the two subsets. It reveals that fake news in the dataset generally has a longer title than that of real news. The overall domain of the length of the title of fake news is also much wider.

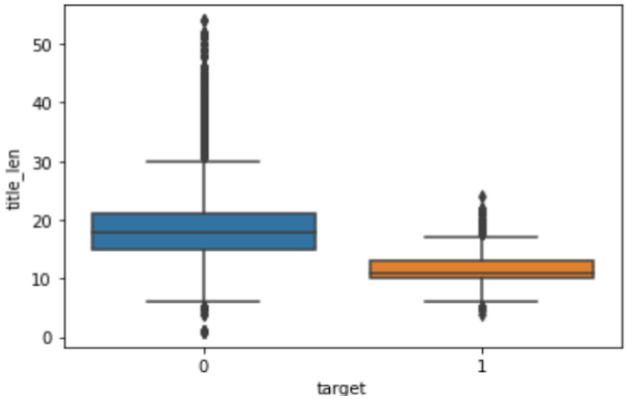

Figure 5. Box plot of lengths of "title" attributes in Dataset 1.

Lastly, the examination of the attribute: 'text', is done in four different approaches: numbers of tokens, sentences, capital letters, and punctuation signs. Figure 6. and 7. tell that the number of tokens and that of sentences of both fake and real news are more or less the same. Thus, these fields are not discriminative for our problem and will not be considered in the future. In contrast, the two boxplots in Figure 8 and Figure 9 signal apparent differences in both capital letters and punctuation signs between the two subsets. Both numbers are much higher in the subset of fake news than in that of real news. This discovery is crucial as it is viable to leverage this information to improve the classification of the research further.

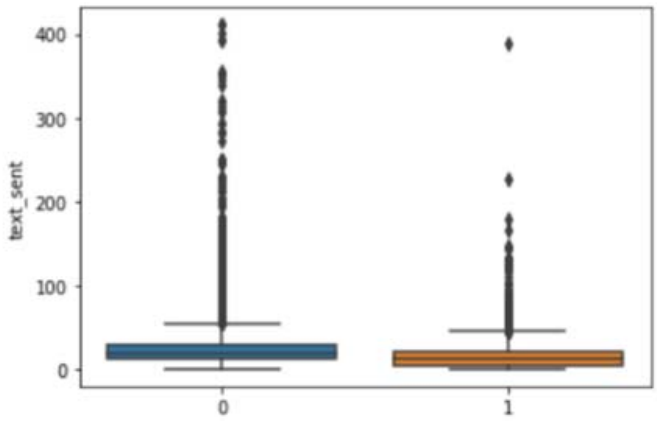

Figure 6. Boxplot of numbers of tokens in Dataset 1.

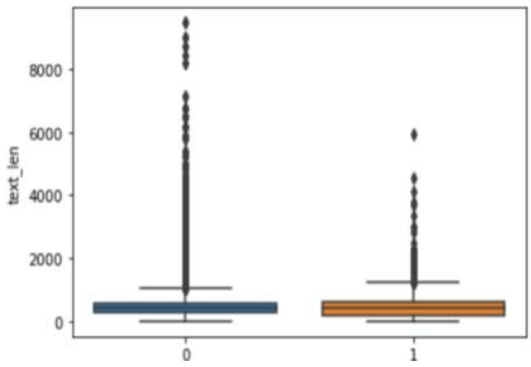

Figure 7. Boxplot of numbers of sentences in Dataset 1.

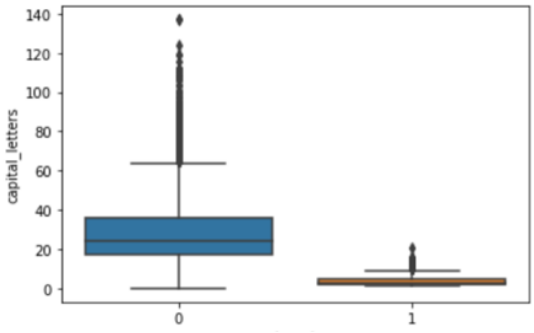

Figure 8. Boxplot of numbers of capital letters in Dataset 1**.**

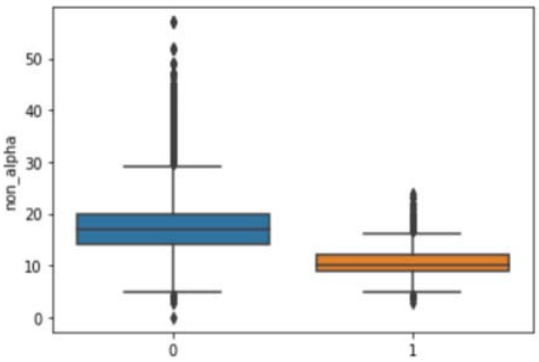

Figure 9. Boxplot of numbers of punctuation signs in Dataset 1.

(2) The EDA of Dataset 2 follows a similar pattern of that of Dataset 1. Firstly, the WordClooud is created for real news and fake news in Figure 10 and 11.

Figure 10. WordCloud of Valid Job Postings.

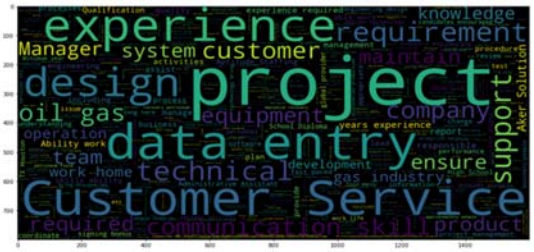

Figure 11. WordCloud of Fraud Job Postings.

Secondly, because of the large number of features in Dataset 2, it is crucial to simplify the dataset by filtering out non-significant information. A heatmap for the correlations between those numeric features is drawn to eliminate useless features. As shown in Figure 12, both of the two attributes, "job_id" and 'telecommuting,' have low correlations with every other attribute, thus are safely dropped.

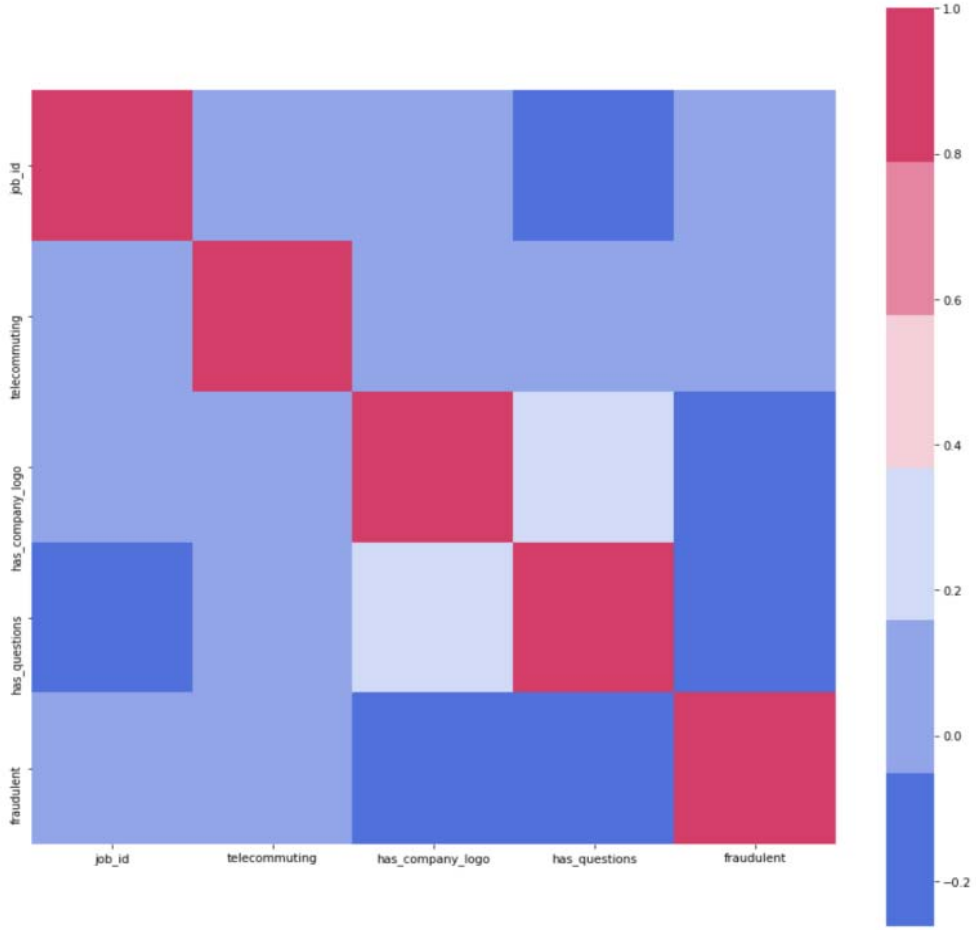

Figure 12. Heatmap of correlations between numeric features in Dataset 2.

3.3. Data Cleaning and PreProcessing

Based on the results of EDA in part B, both Dataset 1 and Dataset 2 are cleaned and processed. The following are the methods been used:

1. *Delete unrelated attributes according to the correlation coefficient*
2. *Fill NAN values with an empty string.*
3. *Remove stopwords, punctuations [12] [13]*
4. *Lemmatizing [14] [15]*

3.4. Vectorization

After the datasets are cleaned, the words and sentences were converted into numerical data through the vectorization algorithms. For each model, a suitable algorithm is used for preprocessing.

3.5. Models

Five of the most prominent models are selected in this work, Long Short-Term Memory (LSTM), Multinomial Naïve Bayes (MNB), Gaussian Naïve Bayes (GNB), Random Forest (RF), and Logistic Regression (LR), respectively.

4. Experiments

4.1. Datasets

For dataset 1, after EDA and data preprocessing, each news in dataset 1 contains two features, 'contents', which contain title and text, and 'class', which represent the real/fake news. As for datasets 2, as what can be seen in Figure 2, the dataset is extremely imbalanced. Thus, the Sampling method is used to make the dataset more balanced. Here, the Over-Sampling method—Synthetic Minority Oversampling Technique (SMOTE)—is used to deal with the imbalanced datasets. The method of synthesizing data was based on KNN, which K equals 5. The parameter is set only to synthesize the minority class, and the distribution of each class after SMOTE is shown in Figure 13.

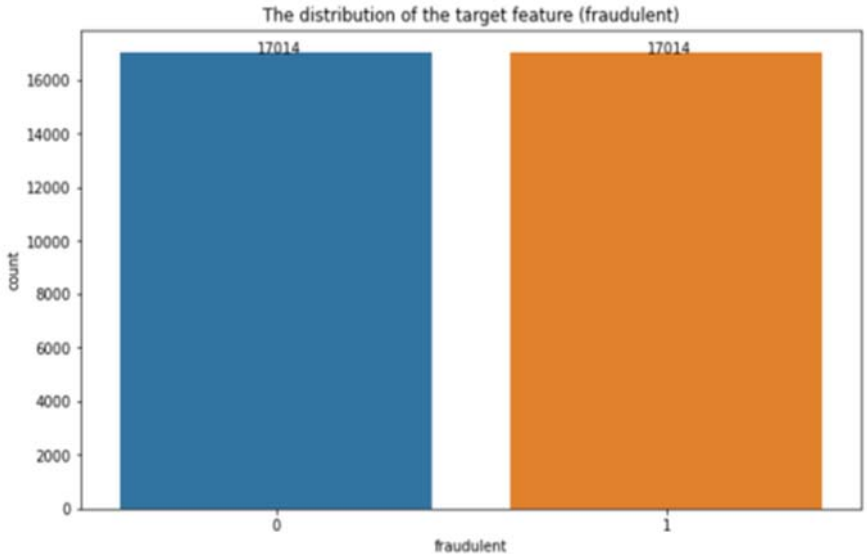

Figure 13. Distribution of negative and positive class.

Finally, dataset 2 contains two features, 'content' and 'fraudulent,' respectively.

4.2. Evaluation Criteria

To avoid one-sided reflection of the performance of the model, selected multiple evaluation criteria can be used to evaluate the model.

- *Precision Score[16]*

In the paper, TP is used to represent True Positive, TN to represent True Negative, FP to represent False Positive and FN to represent False Negative.

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

However, Precision Score can not tell the whole story. It does not indicate how many real positive class examples were predicted as belonged to the negative class, so-called false negatives. So Recall Score is introduced.

- *Recall Score[17]*

Recall provides an indication of missed positive predictions. In this way, Recall provides some notion of the coverage of the positive class.

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

- *f1-score[18]*

For imbalanced datasets, as mentioned earlier, maximizing precision will minimize the number false positives, whereas maximizing the Recall will minimize the number of false negatives. Thus F-measure is introduced to provide a way to express both concerns with a single score.

$$F - Measure = \frac{(2*Precision*Recall)}{(Precision+Recall)} \quad (3)$$

- *Sensitivity/Specificity*

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (4)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (5)$$

- *Roc-Auc*

ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

4.3. Results and Comparison

As stated in Part 3, five models are chosen to classify the text, and the following are the results. The results of dataset1 and dataset 2 are shown in Table 1 and Table 2, respectively.

4.3.1. Dataset 1—Real/Fake News dataset

Table 1. Results of Dataset 1.

Model	Precision	Recall	Sensitivity	Specificity	F1-score	Roc-auc	Accuracy
LSTM	0.9879	0.991	0.991	0.9892	0.9895	0.9992	0.9901
GNB	0.9222	0.9323	0.9323	0.9288	0.9272	0.9605	0.9305
MNB	0.9225	0.9296	0.9296	0.9293	0.9261	0.9802	0.9295
RF	0.9973	0.996	0.996	0.9976	0.9967	0.9998	0.9968
LOG-REG	0.9940	0.9903	0.9903	0.995	0.9921	0.9993	0.9928

Conclusion:

From Table 1, the best-performed model is Random Forest (RF), the recall score of RF can reach up to 0.996. Also, the execution time of RF has much shorted than LSTM. Also, the performance of the Naïve Bayes (NB) model is below average. However, on the bright side, the execution times of NB models are the shortest among five models, so NB is the least time-consuming model. LSTM is intermediate among all models. However, LSTM is the most time-consuming mode when training. After the training and testing process, the confusion matrix is visualized to help us have a clear view of TP and TN rates. In the testing process, the original dataset is split into training datasets and test datasets, the ratio of the training set to test set is 3:1. To sum up, RF, LSTM, and Log-Reg perform roughly the same on dataset 1. On the contrary, the performance of NB models is below average, but their training time is shorter than the other three models.

4.3.2. Dataset 2—Real/Fraud Job Postings

Table 2. Results of Dataset 2.

	Precision	Recall	Sensitivity	Specificity	F1-score	Roc-auc	Accuracy
LSTM	0.9813	1	1	0.9809	0.9905	0.9992	0.9904
GNB	0.8872	0.9953	0.9953	0.873	0.9381	0.9348	0.9342
MNB	0.899	0.8939	0.8939	0.8991	0.8964	0.9681	0.8965
RF	0.9756	0.9967	0.9967	0.975	0.986	0.9986	0.9858
LOG-REG	0.9893	1	1	0.9885	0.9946	0.9995	0.9944

Conclusion:

For the performance of models on dataset 2, things are a bit different since the dataset is exceptionally imbalanced. To deal with the imbalanced dataset and avoid over-fitting, Over-Sampling technique (SMOTE) was implemented, after sampling, the number of negative classes is the same as the number of positive classes, as shown in Figure 13. From Table 2, the best-performed model is Log-Reg as its recall score can reach up to 1.0. LSTM is also great even though its training time is relatively longer than the other four models. RF does not perform well as on dataset2, because there is still some noisy data after SMOTE and over-fitting happened. For Bayes models, the results are the same as dataset 1. Their performances are below average. However, they are much less time-consuming, and it is a good choice when practical application mainly focuses on training time.

4.4. Text visualization

In addition to simply train and test model in two datasets, our research also visualizes the text content of dataset 1, in detail, Most 20 frequent words in real news are shown in Figure 14, Most 50 important features are shown in Figure 15. This is practical since it can help people distinguish real or fake news according to the words in the news content.

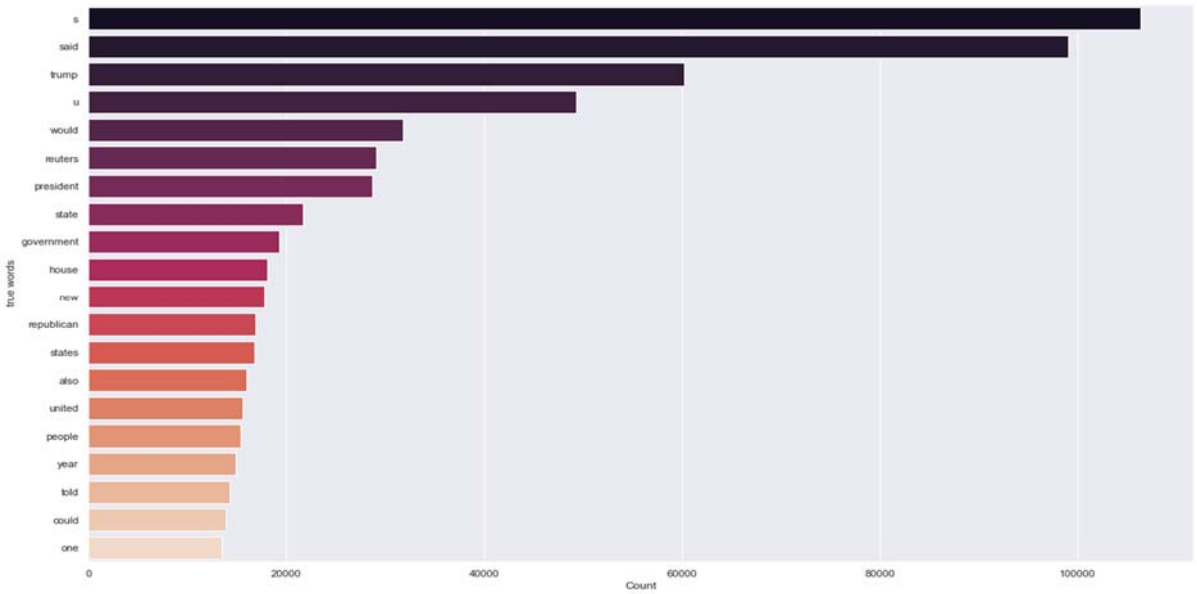

Figure 14. Most 20 Frequent words in Real News.

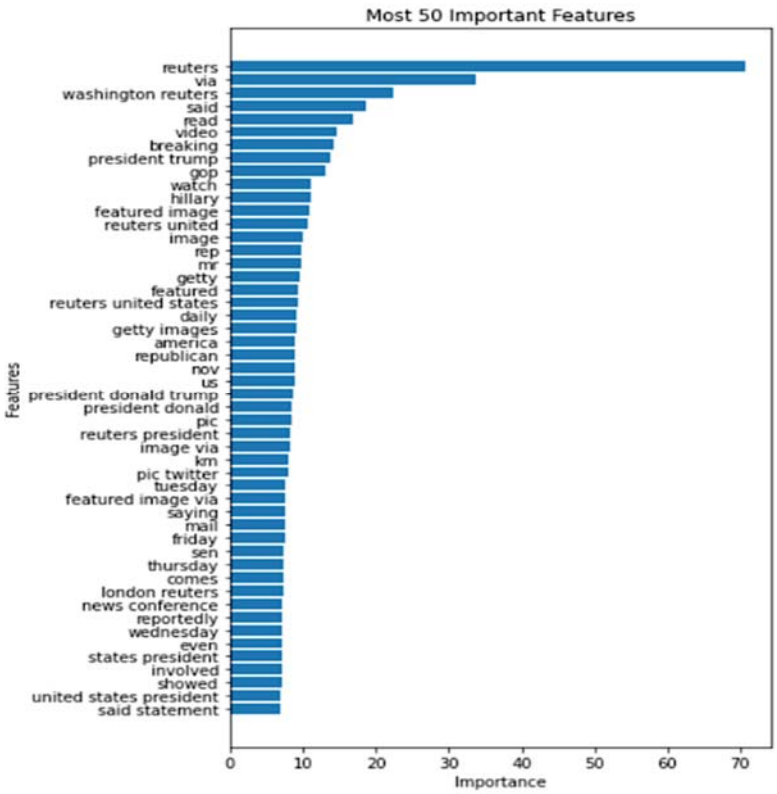

Figure 15. Most 50 important words.

5. Conclusion

By executing Exploratory Data Analysis on both two datasets, we are able to clean up both Dataset 1 and Dataset 2 a lot in many ways. The over-sampling technique helps us deal with the imbalanced dataset; thus, balance weights of both classes and avoid over-fitting.

The usefulness of a fake news detection model depends much on its accuracy, however, at the same time, efficiency is also crucial because of the temporal nature of the news. Combining the results of experiments using both datasets, Naïve Bayes is the most time-efficient model but with the lowest accuracy performance in general. While the rest all perform similarly on balanced datasets, Logistic Regression is the winner on Dataset2, while RF being the most vulnerable among the three. LSTM is shown to be a very decent choice in classifying fake news; however, the long training time may pull it back some time.

As for future work, firstly, more researches are planned to do on vectorization algorithms since various algorithms may have different effects on the same model. Moreover, our datasets are mainly about news and postings from the Internet, so these models are planned to apply into different topics and fields, such as shopping experiences from online websites, comments of books and films from social media, etc. Lastly, imbalanced datasets are common in real-world problems, and more effective techniques are ready to be found to solve the problem.

6. Contribution

In this work, Peiyang conceived the project, wrote most of the coding work, Victor finished Vectorization part of the coding work, and Jiaxin finished Naïve Bayes coding work. Peiyang, Victor and Jiaxin finished writing work together. All authors read and approved the final manuscript.

References

[1] https://www.lexology.com/library/detail.aspx?g=ab8f79df-3bfe-4f01-8cee-b85b627f6226

[2] https://www.hstoday.us/subject-matter-areas/emergency-preparedness/fake-news-during-disasters-putting-peoples-lives-at-risk-warns-intel-bulletin/
[3] https://pursuit.unimelb.edu.au/articles/fake-news-in-the-age-of-covid-19
[4] J. C. S. Reis, A. Correia, F. Murai, A. Veloso and F. Benevenuto, "Supervised Learning for Fake News Detection," in IEEE Intelligent Systems, vol. 34, no. 2, pp. 76-81, March-April 2019, doi: 10.1109/MIS.2019.2899143.
[5] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston, Journal of Chemical Information and Computer Sciences 2003 43 (6), 1947-1958
[6] Huang, Jie; Zhou, Wengang; Zhang, Qilin; Li, Houqiang; Li, Weiping (2018-01-30). "Video-based Sign Language Recognition without Temporal Segmentation".
[7] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. In Proceedings of the 26th ACM international conference on Multimedia (MM '18). Association for Computing Machinery, New York, NY, USA, 117–125. DOI:https://doi.org/10.1145/3240508.3240533
[8] Has¸im Sak, Andrew Senior, Franc¸oise Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," Google, USA
[9] Rymarczyk T., Kozłowski E., Kłosowski G., Niderla K. Logistic Regression for Machine Learning in Process Tomography. Sensors. 2019;19:3400. doi: 10.3390/s19153400.
[10] https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
[11] https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction
[12] https://towardsdatascience.com/stop-words-in-nlp-5b248dadad47
[13] https://www.quora.com/Why-should-punctuation-be-removed-in-Word2vec
[14] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
[15] https://www.datacamp.com/community/tutorials/stemming-lemmatization-python
[16] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
[17] http://datascience.sharerecipe.net/2020/01/02/how-to-calculate-precision-recall-and-f-measure-for-imbalanced-classification/
[18] https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/