Group-7: Fake News Detection

Mihir Shah 1150679
Kavish Narula 1152238
Karan Patel 1148274
Nishit Shah 1149657
Pearly Shah 1147841

Abstract

The proliferation of fake news during the 2016 Presidential election in United States highlighted not only the dangers of the consequences of fake news but also the difficulties presented in separating fake news from real news. Fake news for different economic and political goals has been surfacing in huge numbers and spreading in the internet world. Detecting fake news is a crucial task, which not only entails users receive original verified information but also helps maintain a trustworthy news ecosystem. The major portion of current detection algorithms target on finding clues from news contents, which are generally not effective because fake news is often intentionally written to mislead users by mimicking true news. The results of this project demonstrate the ability of Natural Language Processing to be useful in this task. We have built a system that catches many intuitive indications of real and fake news comparing performance of various models such as SVM, Random Forest, CNN, LSTM, logistic regression and Multinomial Naive Bayes geared with TFID word to vector and LeMA.

1 Introduction

Misinformation is a kind of foreign press that deliberately promotes inaccuracies or frauds through forecasting mainstream media including more recent online networks. Fake news, willful falsehoods, scams, mimics, and satire are all examples of how to deceive people in order to harm an institution, organization, or person, and benefit politically or financially. Because of how it might affect the present financial situation, false stories have been in the attention of media organizations and the community at large scale. Social media seems to be the powerful source of circulating such information, and it occasionally makes its way into the mainstream media as well.

Since the release of the ”Great moon hoax” in 1835, there’s been false information(2). Fake News are pseudo-scientific and usually diminishes the publication of real headlines. The project’s primary goal is to detect and classify false news. Fraudulent eye-catching and interesting claims are made to encourage viewers’ attention in try to acquire material. This is especially troublesome and is used as a weapon to end corruption. As a result, the challenge of detecting false news requires far more attention than it presently receives. The major difficulty in resolving the issue of fake news is the ambiguous definition of the term. Fake news, for example, may be classified into several categories: a remark that is known to be entirely untrue, or a speech that presents numbers as facts without any actual analysis. TFID vector and LeMA have been introduced along with the models to further accurately classify fake news statements. Here TFID comprises of Term frequency which means the number of times a word appears in a document divided by the total number of words in the document and Inverse Document Frequency which results in log of the number of documents divided by the number of documents that contains word ’w’. Moreover, Lemmatization(LeMA) is the process to convert the word of token to its base form, for example. House, Houses and Housing will be converted to its base form that is House. Several attempts have been made, but we really do not have a viable method for detecting false news. Oliveira and Figure (2017).

2 Problem Statement

In Traditional news circulation process, content flow was controlled, censored and non-amplified. Whereas, news circulation process in the internet age has introduced many problems related to fake circulation, uncensored and amplified news as anyone can edit and publish the data. The purpose of the project is to classify the news with the help of NLP models into categories based on level of truth and falsity in each statement.

3 Related Work

In this project we are gaining the knowledge by reading this research work and building a new custom structure by cross validate the model and also by converting statement with the unique id feature to analysis and to accurate the model performance using Random forest, Gaussian Naive Bayes, SVM and Logistic regression.

- M.Mokhtar et al. in the paper(11) fakebuster proposes fake news detection using Logistic Regression Model.Based on the study model showed a good performance in classification task.Based on analysis made, logistic model within stance detection approach yields an excellent accuracy using

TF-IDF feature in constructing this fake news model. This model when integrated with web services accepts input either news URL or news content in text which is then checked for its truth level through “FAKEBUSTER” application.As a result, by using n-gram for content based the accuracy obtained is 80% and for Stance based the accuracy is 99%.

- S.Yousuf et al. in the paper (4), proposed two different machine learning algorithms i.e random forest and decision tree to detect the fake news. In this research paper, the size of dataset contains 20,761 samples, while the testing sample size equals 4,345 samples. The preprocessing of dataset is done which includes cleaning data by removing unnecessary special characters, numbers, English letters, and white spaces, and finally, removing stop words is implemented. After Preprocessing, TFIDF feature extraction method is used before applying the two suggested classification algorithms. As a result the best accuracy achieved is 89.11% using the decision tree model while using the random forest; the accuracy achieved is 84.97%.
- K.Stahl et al. in the paper(12), they have considered past and current techniques for fake news identification in text formats while elucidating how and why news fake exists in any case. This paper incorporates a discussion on how the writing style of a paper can also impact on its classification. They had implemented their project using Naïve Bayes Classifier and Support Vector Machines methods. They had looked into the semantic analysis of the text for classification.
- D. Elisabeth et al in the paper (10) describes the performance of hate code detection for Indonesian tweets using machine learning and a classification explainer. Two scenarios are taken into consideration when performing the project i.e hate code from hate speech classification and hate code from hate code classification.The models used are random forest decision tree ,Logistic regressions, Naive Bayes. For tweets that annotated have hate code, the f-measure is 28.23% for recognized all the hate codes, and the recall is 56.91%.The best f-measure score is 94.90% from hate code classification using Logistic Regression with abusive codes elimination.

4 Motivation

Advancement in technology will keep on happening along with the time and so the problem that we are seeing today can become tomorrow’s big issue of soceity. People have started using social media in a wrong way and one small fake news can make or break a company. For instance it reminds me of a situation when a company named ”Adani Group” suffered from a big loss because one of it’s competitor had spread a rumour about the company. And in no time prices of this stock plunged drastically. This did not impact the big investors but definetly took all the savings from common people who had trust in this company and who kept all their hard earned money into it. There are some other examples also where fake news has done enough damage to mankind which needs immediate attention and thus to do our part of the soceity we thought of giving our best to solve this problem.

5 Methodology

As per the introduction there are plenty of techniques nowadays for detecting fake news and also the old method are less accurate with time consuming due to legal action that can be done with the help of different methods. Method involved after the involution of modern era with various deep learning technology and models such as logistic regression, CNN, LSTM, RES-NET, SVM, Random forest and others. In this project we have firstly done with the implementation of CNN and Bi-LSTM before our progress for the initial stage of project to check the accuracy of fake news detection.

5.1 Approach

After facing different problem and have very low accuracy which is less than 30 percent in both Bi-LSTM and CNN then we moved forward and rebuild our model structure import new model such as SVM, Random forest, logistic regression and Multinomial Naive Bayes having conversion of statement into TFID, converting statement features using WordToVec of the extracted feature of statements and by Lemmatisation for to get better result from previous one.

5.1.1 Logistic regression

A logistic regression (LR) model is used to classify text based on a large feature set with a binary output (true/false or true article/fake article), since it gives an easy equation to classify issues into binary or many classes (7). Our tuned hyper parameters to achieve the best results for each dataset individually, whereas several values were evaluated first. The output of logistic regression is processed into a confidence interval using a Sigmod function; the aim is to limit the cost function to get the best possibility.

5.1.2 Support Vector Machine(SVM)

Another model for binary classification problems is the support vector machine (SVM), which is available in a variety of learning algorithm (6). An SVM model’s goal is to characterize data sets by constructing a subspace (or support vectors) based on a set of features (8). The size of the vector space is calculated by the size of attributes. Since one centroid potentially reside in countless locations in an N-dimensional space, the problem is to discover the plane that separates the data points of two classes with the validates that will having True and false as its final presentation.

5.1.3 Multinomial Naive Bayes

Naive Bayes Classifier Algorithm is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of a feature. Bayes theorem calculates probability P(c—x) where c is the class of the possible outcomes and x is the given instance which has to be classified, representing some certain features(9).

P(c—x) = P(x—c) * P(c) / P(x)

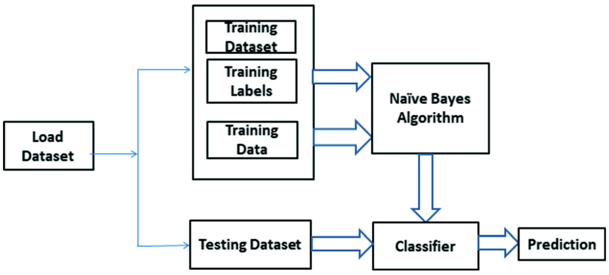


Figure 1: Multinomial Naive Bayes Model

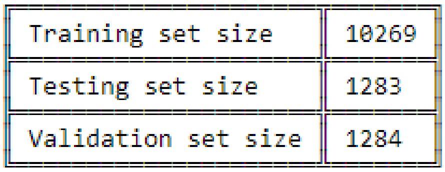
Naive Bayes are mostly used in natural language processing (NLP) problems. Naive Bayes predict the tag of a text. They calculate the probability of each tag for a given text and then output the tag with the highest one.

5.1.4 Random Forest

Random forest (RF) is a supervised learning method that is an upgrade of decision trees (DT). RF is completely invented of a significant number of decision trees that work together to predict the outcome of a class, with the final prediction based on the class that gained the most votes. Due to the relatively low connection among trees, the error rate in random forest is negligible when compared to other methods (3). Our random forest model was trained with different parameters, such as various amounts of explanatory variables, in a proposed method to find the optimal model that can accurately predict the future.

5.2 Dataset

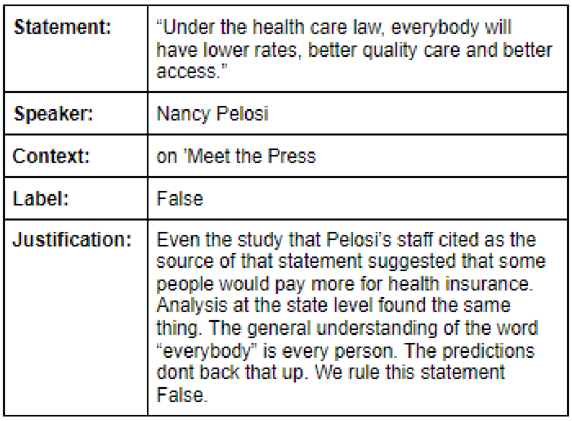
In this project we have used publically available dataset named LIAR (1) dataset by Wang for Fake news detection.Here decade long data was collected, it contains 12,836 manually labeled short statements from POLITIFACT.COM. The statement are labeled for truthfulness, subject, context/venue, speaker, state, party, and prior history. The instances collected in LIAR grounded, more natural context, such as Facebook posts, political debate, tweets, interview, TV ads, news release, etc. In each case, the labeler provides a lengthy analysis report to ground each judgment, and the links to all supporting documents are also provided.The below table shows the statistics of the dataset.



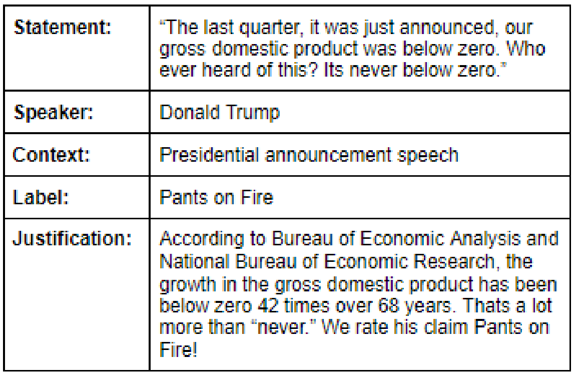
Training set size	10269
Testing set size	1283
Validation set size	1284

Table 1: Dataset Statistics

Below is some random excerpts from the LIAR dataset.



Statement:	"Under the health care law, everybody will have lower rates, better quality care and better access."
Speaker:	Nancy Pelosi
Context:	on 'Meet the Press
Label:	False
Justification:	Even the study that Pelosi's staff cited as the source of that statement suggested that some people would pay more for health insurance. Analysis at the state level found the same thing. The general understanding of the word "everybody" is every person. The predictions dont back that up. We rule this statement False.



Statement:	"The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero."
Speaker:	Donald Trump
Context:	Presidential announcement speech
Label:	Pants on Fire
Justification:	According to Bureau of Economic Analysis and National Bureau of Economic Research, the growth in the gross domestic product has been below zero 42 times over 68 years. Thats a lot more than "never." We rate his claim Pants on Fire!

Table 2: Glimpse of dataset

5.3 Evaluation

The main objective of this project is to predict the authenticity of news, And of that we are using Liar Dataset. The dataset comprises of 12.8k manually labeled short headlines/statements collected over the decade in various contexts from "PolitiFact.com".

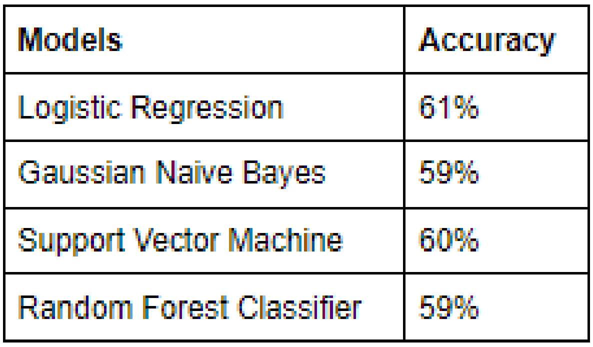
To further process the dataset, We examined the distribution of labels and came to a conclusion that labels were uniformly distributed. Moreover, we will classify the test data based on class predicted for the news statement. Lastly, we plan to compare several results by tuning hyper-parameters and conclude output with the best fit.

The next phase to the model is to run the architecture using pipeline for to get best accuracy by cross validate the classifier and also by converting the extracting feature with WordToVec format for analysing the problem to detect fake News from the speaker's point of view by LIAR dataset.

Further more by running the model using TFID, WordToVec and Grid-Search view the accuracy we are not getting that much as require for the fake news detection. So by we have used Lemmatisation for the model accuracy in that we are successfully getting approx 59% to 61% accuracy for the project that is better to detect weather the news are fake or not.

5.4 Result

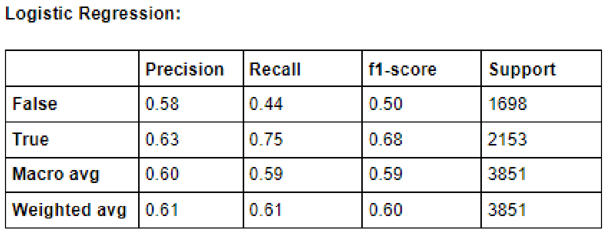
We have used 70% training dataset and 30% of testing dataset for the evaluation of our project.The below table shows the accuracy performed by each model.



Models	Accuracy
Logistic Regression	61%
Gaussian Naive Bayes	59%
Support Vector Machine	60%
Random Forest Classifier	59%

Table 3: Result

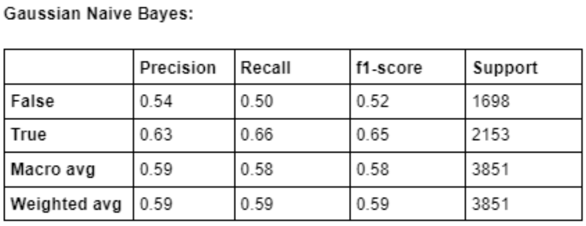
5.4.1 Performance of Individual Models



Logistic Regression:

	Precision	Recall	f1-score	Support
False	0.58	0.44	0.50	1698
True	0.63	0.75	0.68	2153
Macro avg	0.60	0.59	0.59	3851
Weighted avg	0.61	0.61	0.60	3851

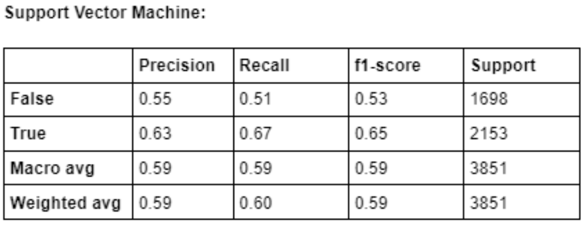
Table 4: Logistic Regression Evaluation



Gaussian Naive Bayes:

	Precision	Recall	f1-score	Support
False	0.54	0.50	0.52	1698
True	0.63	0.66	0.65	2153
Macro avg	0.59	0.58	0.58	3851
Weighted avg	0.59	0.59	0.59	3851

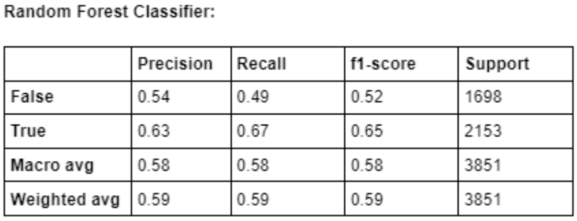
Table 5: Gaussian Naive Bayes Evaluation



Support Vector Machine:

	Precision	Recall	f1-score	Support
False	0.55	0.51	0.53	1698
True	0.63	0.67	0.65	2153
Macro avg	0.59	0.59	0.59	3851
Weighted avg	0.59	0.60	0.59	3851

Table 6: Support Vector Machine Evaluation



Random Forest Classifier:

	Precision	Recall	f1-score	Support
False	0.54	0.49	0.52	1698
True	0.63	0.67	0.65	2153
Macro avg	0.58	0.58	0.58	3851
Weighted avg	0.59	0.59	0.59	3851

Table 7: Random Forest Classifier Evaluation

6 Future Scope

So from the experimental result and applying different model to analysing the Natural language processing by using different conversion of the the statement for extracting the feature from the dataset and using them as unique ID to test the data for the accuracy. But by using the lemmatisation method to detecting the false statement of the speaker of their respective parties. Our future goal is to work on different dataset that will related to this fake news such as movie sentiment analysis, movie review, recommendation of different things to buy the product and other using this TFID, WordToVec and Lemmatisation method. Also we will using the deep learning model such as RESNET, ALexNet and other various model to compare the accuracy. Also our main is to built a system for the social media application to Erase and not to post any false news to the platform with the help of this model and the data structuring we will build a fully secure and NLP based Fake news detection for social media platform as nowadays it is the fastest and quick spreading platform for any type of news.

7 Conclusion

As clearly visible in Fig. 1, All the algorithms performed almost equally(around 23 percent accuracy) with custom basic setup. Furthermore, When introduced with TFID(Term Frequency and Inverse Document Frequency), there was a minute increase in accuracy for all algorithms. Finally to improvise performances, we introduced Lemmatization. The performances of Random Forest, Naive Bayes and SVM surged drastically and more than doubled. Amongst all, SVM with LeMA outperformed giving an accuracy of 61 percent.

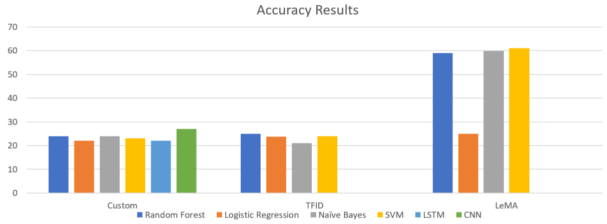


Figure 2: Model Performances

References

[1] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

[2] Great moon hoax. https://en.wikipedia.org/wiki/Great Moon Hoax. [Online; accessed 25-September-2017]

[3] "Approaches to Identify Fake News: A Systematic Literature Review.", Dylan de Beer and Machdel Matthee. 2021.

[4] Al-Shammari, Reham Yousif, Suhad A. (2020). Fake News Classification Using Random Forest and Decision Tree (J48). 23. 8.

[5] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," Statistics and Computing, vol. 27, no. 3, pp. 659–678, 2017.

[6] "N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.

[7] "T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," The Annals of Statistics, vol. 36, no. 3, pp. 1171–1220, 2008. View at: Publisher Site — Google Scholar

[8] "T. M. Mitchell, The Discipline of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA, 2006.

[9] "Applying Multinomial Naive Bayes to NLP Problems", https://www.geeksforgeeks.org/applying-multinomial-naive-bayes-to-nlp-problems/, 14 Jan 2019

[10] D. Elisabeth, I. Budi and M. O. Ibrohim, "Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study," 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1-6, doi: 10.1109/ICoICT49345.2020.9166251.

[11] M.Mokhtar, Y.Jusoh, N.Admodisastro, N.Pa,A.Amruddin,"Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning",2019, https://doi.org/10.35940/ijeat.a2633.109119

[12] K.Stahl., "Fake news detection in social media", B.S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, 2018.

8 Appendix

- List of Figures:

- List of Tables:

- List of Libraries
  - ftfy:
    It fixes Unicode that's broken in various ways.
  - nltk:
    NLTK consists of the most common algorithms such as part-of-speech tagging, stemming,tokenizing,topic segmentation, sentiment analysis, and named entity recognition.
    - nltk.stem
    - nltk.tokenize
    - nltk.corpus
  - seaborn
    Seaborn is a Python data visualization library based on matplotlib.
  - pandas
  - numpy
  - matplotlib
    Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
  - sklearn
    The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction
    - sklearn.model_selection
    - sklearn.base
    - sklearn.pipeline
    - sklearn.feature_extraction
    - sklearn.feature_selection
    - sklearn.linear_model
    - sklearn.naive_bayes
    - sklearn.preprocessing
    - sklearn.ensemble

- Software Used:
  Google Colaboratory

- Contribution:
  Each and every member of group 7 have worked under every task given by the professor for the Project i.e from Project Selection to the final report.Whenever there is new task updated we had done the meeting and divided the whole task into five task for each group members and contributed our best in the Project.Below are the list of work done by the group members:

- Mihir Shah:
  Analyzing problems and implementing different methodology; Support Vector Machine and Logistic Regression with LeMA and TFID vector.
- Kavish Narula:
  Analyzing problems and implementation of counter vector for extracting statements and Report Formatting.
- Karan Patel:
  Dataset Preprocessing, Literature Review, Implementation of TFID, Documentation and Presentation.
- Nishit Shah:
  Analyzing problems and implementing different methodology; mainly Multinomial Naive Bayes and Random Forest.
- Pearly Shah:
  Data searching, Dataset Execution and normalization, implementation of Gaussian Naive Bayes and all other models without TFID.