Fake News Detection

Jinish Trivedi
Department of Computer Science and Engineering
Institute of Technology
Nirma University
Ahmedabad, India
20bce109@nirmauni.ac.in

Malay Jiyani
Department of Computer Science and Engineering
Institute of Technology
Nirma University
Ahmedabad, India
20bce111@nirmauni.ac.in

Kashish Kakadiya
Department of Computer Science and Engineering
Institute of Technology
Nirma University
Ahmedabad, India
20bce121@nirmauni.ac.in

Kishan Gondaliya
Department of Computer Science and Engineering
Institute of Technology
Nirma University
Ahmedabad, India
21bce511@nirmauni.ac.in

***Abstract*—People learn how to discriminate between real and fraudulent information via experience and aging. Children are trained to recognize falsehoods and dishonesty from an early age. People have also been able to detect lying in general for decades. The only reliable sources of information nowadays are television, newspapers, and radio. So why would people are waiting a whole day for the newspaper to publish or several hrs for the news tonight when they can simply click on Facebook and read their feed? After all, it is the reason people follow news organizations' Facebook pages. A poll of 137 people, including students in high school college students, and graduates, as well as people working in a range of professions including engineers, lawyers, educators, librarians, designers, etc., was done. The purpose of the study is to look at how well social network users can spot clickbait and bogus news. When creating models using machine learning for false news categorization, education programs, and solutions for false information detection and prevention, such research may be helpful. To find out the overall population's degree of information literacy, a larger-scale study like to this one may be carried out.**

I. Introduction

The term "Fake News" refers to the contact that can be verified with respect to the original content. Nowadays fake news spread like a virus in the system, which affects everything that is attached to it. The main reason behind this widespread is social media, social media is the biggest reason for the growing phase of fake news because it is a platform where people are highly active. Due to highly activation of users in social media, this platform totally misleads information related to the actual content. Social media has gained so much popularity and attention that people express their ideas and opinions publicly and the public are like to obtain news that is false/fake due to this misleading platform. Fake news has proved that it is highly toxic and feckless for communities like businesses, education, the government sector, etc. It is necessary to stop the dissemination of fake information or false information and circulation take place due to the vast volume of false data that cannot be processed by actual content checking.

We can create a classifier dataset of actual news and fake news that can make a proper judgment about information. We can predict that if a source is labeled and detected as a contributor to fake news then we can predict that all the sources from this label are illegal and completely fake. Using machine learning we can create two models and compare them to check the more accurate possibility. Concepts like Random forests and Decision trees are used to construct a model for the classification and detection of fake news. The experiments performed on the dataset utilized in this paper appear to have accurate findings. We were inspired to pursue this project by the expanding trend in this field. The DT model is used to categorize fake news in a novel way. In order to describe the course of action in the decision tree, the proactive personalities are calculated using a tree-shaped pattern. The best decision tree is not built by RF algorithms when it comes to comparison. The primary objective of the framework provided is to take into account the significance of form, and semantic similarity, and create DT algorithms to classify the proactive personality data, detect fake news, and prevent it from reaching the public.

II. Methodology

A. Dataset

In the proposed study, the data set is acquired, the data set is also used, and machine learning techniques are applied. The fake news dataset was downloaded from the Kaggle website. The data consists of 44898 entities with 5 attributes namely Title, Text, Subject, Date, and Target.

B. Data Visualization

The above image shows a comparison between fake and true news that is spread on social media like a virus. We used

Identify applicable funding agency here. If none, delete this.

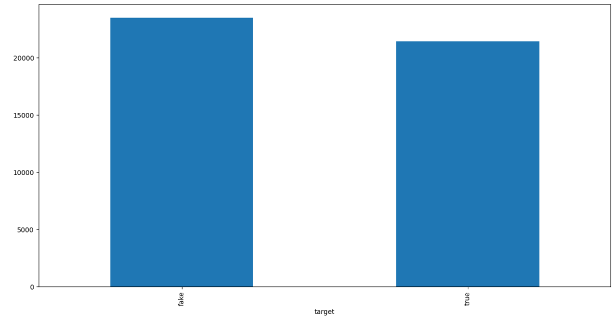


Fig. 1. No of Fake and True news

a count function to calculate the number of counts fake and true, and accordingly, we plotted a graph.

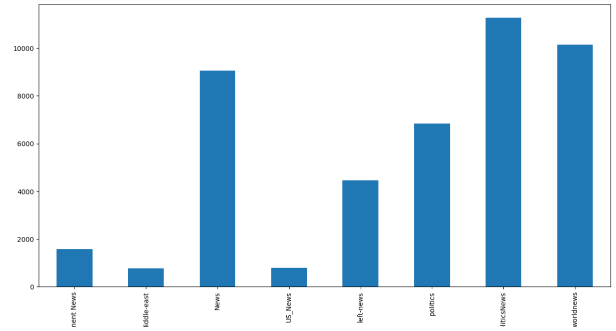


Fig. 2. Subject wise no of news

This image represented the most common topics widely used by a news channel, Whenever any fake news is spread these are some of the topics which are widely observed and these are some fresh topics that are influenced or targeted by fake news.

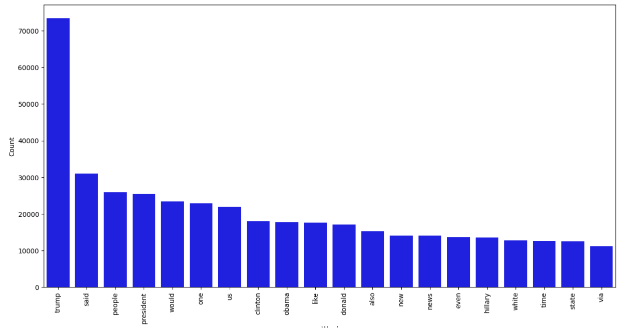


Fig. 3. Count most frequent word in fake news

This image represented the most common fake news words that are used in articles, and messages(social media). We have implemented a function that stores and count the occurrences of these fake word to identify whether the news is fake or true.

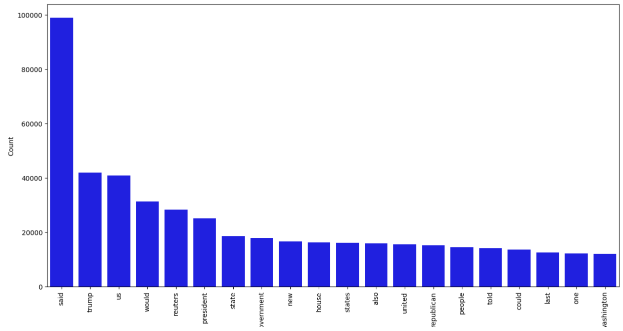


Fig. 4. Count most frequent word in true news

This image represents the most common word used in fake news. By storing these words we can identify whether the news is fake or true because these words are stored in a counter function.

C. Data Preprocessing

1. Title
2. Text
3. Subject
4. Date
5. Target

Dataset has a total 44898 number of data. Every dataset has some unrelated features and redundancy, which leads to wrong predictions, less accuracy, and more errors. We need to remove it that process is called data pre-processing. We need to clean the dataset before the building process to improve model accuracy. We converted our dataset to a dataframe with only 3 columns.

1. Text: The processed and formatted text.
2. Subject: News related to which subject is important.
3. Target: A ‘fake’ and ‘true’ represent the type of news.
Some features like Title and date are removed and cleaned from the dataset. Some large amount of processing is there for more accuracy and accurate predictions.

1. Punctuation Removal: Regular expression like ”!”&’()*+-,/:;¡=¿?@[]— ” was used for this preprocessing. All these types of specials should be removed for more accuracy.
2. Stop word removal: with the help of nltk library, we have downloaded 17 stop word sets. Removing these stop words improves the model to predict the type of news.
3. To lower case: all the text is converted to lower case, so the upper case will not lead to less accuracy.

D. Feature Extraction

1) CountVectorizer:
Count vectorizer is a very perfect tool provided by the sci-kit-learn library in python. It converts a text into a

vector with the help of the frequency of each word that occurs in the entire text. It is used to convert each word in each text to a vector, it is also helpful and efficient when there are multiple texts. It creates a 2-dimensional matrix in which rows represent each text sample from the document, columns represent each word in each text and each cell represents the count of that particular word in each text.

2) Tf dif Transformer:

Term Frequency- Inverse Document Frequency is the full form of Tf-Dif. It is employed to lessen the influence of tokens that occur frequently in a given corpus, making it experimentally less informative than a characteristic that very sometimes occurs in the training corpus.

E. Classifier Model

1) Random Forest:
It is one of the most powerful algorithms in machine learning under the concept of the supervised learning algorithm. Random forest means a combination of multiple decision trees to create the forest. When we use this concept each tree in the forest gives a classification or “vote”.The forest chooses the classification with the majority of votes therefore using the random forest for regression the first pick is the average of the output of all trees.
Why? Random Forest : Random forest is used on the job by data scientists in many industries including banking, stock trading, medicine, and e-commerce. It’s used to predict the things which help these industries run efficiently, such as customer activity, patient history, and safety.

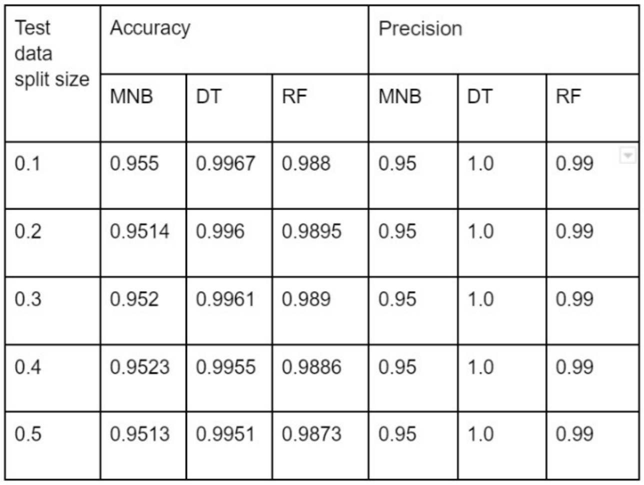
2) Decision Tree:
A supervised learning method called a decision tree may be used to solve classification and regression issues, but it is often favored for doing so. It is a tree-cluster classifier, where internal nodes stand in for a dataset’s characteristics, branches for the decision-making process, and each leaf node for the classification result. It is a visual illustration for obtaining all feasible answers to a choice or problem based on predetermined conditions. It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure like a tree.

3) Multinomial Naive Bayes:
Natural Language Processing uses the Multinomial Naive Bayes algorithm as a probabilistic learning technique most frequently. The method, which guesses the label of a text such as an email or newspaper article, is predicated on the Bayes theorem. For a given sample, it determines the probabilities of each tag and then outputs the label with the greatest chance.
The Naive Bayes classification is a group of many methods, all of which are based on the idea that each feature being categorized is independent of every other feature. The existence or absence of one character has no bearing on the other feature’s existence or absence.

F. Display The Predicted Score

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$PRECISION = \frac{TP}{TP + FP} \tag{2}$$



Test data split size	Accuracy			Precision		
	MNB	DT	RF	MNB	DT	RF
0.1	0.955	0.9967	0.988	0.95	1.0	0.99
0.2	0.9514	0.996	0.9895	0.95	1.0	0.99
0.3	0.952	0.9961	0.989	0.95	1.0	0.99
0.4	0.9523	0.9955	0.9886	0.95	1.0	0.99
0.5	0.9513	0.9951	0.9873	0.95	1.0	0.99

Fig. 5. Accuracy Precision table

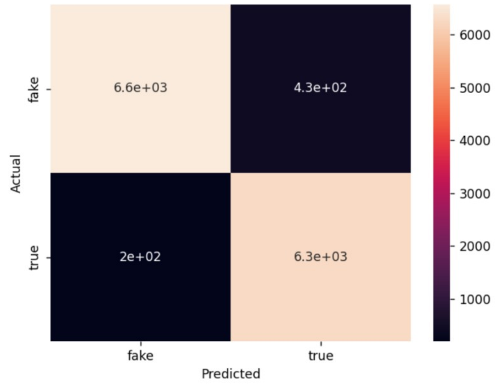


Fig. 6. Confusion matrix for Multinomial

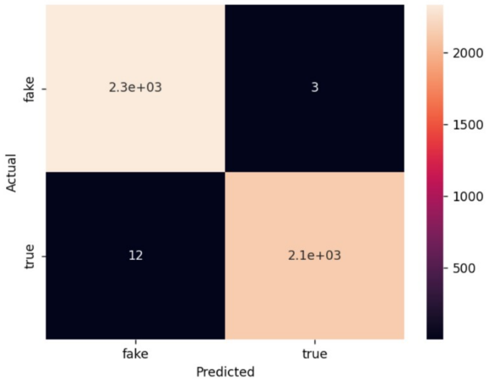


Fig. 7. Confusion matrix for Decision Tree

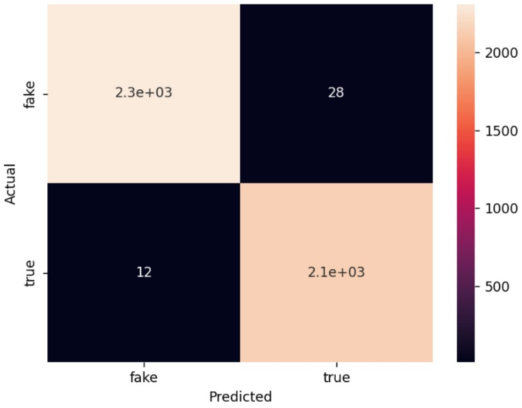


Fig. 8. Confusion matrix for Random Forest

III. Conclusion And Future Work

In this study, we compare the classification of fake and real news using different machine learning algorithms trained on the same datasets. Additionally, we contrasted the outcomes of several feature engineering approaches and their pairings. Additionally, we can use more feature extraction methods, such as Doc2Vec and Glove, which are Deep Learning-based models capable of more effectively learning semantic characteristics. Deep Neural networks and decision trees are two additional Deep Learning classifiers that may be contrasted and analyzed. Another option to consider is experimenting with multimodal data, such as audio and graphics.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.