Fake News Detection Using Machine-Learning Algorithms

Shaneen Shadman Khandakar
Department of Computer Science and Engineering
BRAC University
Dhanmondi, Dhaka
shaneen.shadman.khandakar@g.bracu.ac.bd

Shafkat Zahir
Department of Computer Science and Engineering
BRAC University
Lalmatia, Dhaka
shafkat.zahir@g.bracu.ac.bd

Humaira Tasnim Ridita
Department of Computer Science andEngineering
BRAC University
Gendaria, Dhaka
humaira.tasnim.ridita@g.bracu.ac.bd

Fardin Junayed Karim
Department of Computer Science and Engineering
BRAC University
Uttara, Dhaka
fardin.junayed.karim@g.bracu.ac.bd

Md. Saqif Islam
Department of Computer Science and Engineering
BRAC University
Uttara, Dhaka
md.saqif.islam@g.bracu.ac.bd

Mahin Islam Provat
Department of Computer Science and Engineering
BRAC University
Panthapath, Dhaka
mahin.islam.provat@g.bracu.ac.bd

ABSTRACT

In the current era of technological advancement where mobile phones and other electronic gadgets are becoming more common day by day thus having to deal with the vast availability of information sources on the internet, a genuine challenge arises: to figure out whether the news we are reading online is real or fake. Verifying the authenticity of the news has become quite important for us since it not only helps to eliminate rumors from spreading like wildfire but also prevents misinforming people. This paper's goal is to not only detect fake news but also to find an efficient method and model in the process of the prediction. The fake news prediction is done by cleaning the datasets with preprocessing and then comparing different types of supervised learning and ultimately selecting the one that is the best for the scenario in which case we are achieving an accuracy of 99.67%.

KEYWORDS

Fake News; Supervised Learning; Logistic Regression

I INTRODUCTION

The fast transformation in our communication environment brought about by the internet has resulted in several changes in our society. Social media is usually regarded as the most popular means of daily communication. Social media has had a huge influence on everyone's everyday lives, from youngsters to the elderly. As a result of the widespread use of social media for everything, people are increasingly sharing information without understanding if it is phony or genuine. People are spreading news without first verifying its accuracy. Social media has had a huge influence on everyone's everyday lives, from youngsters to the elderly. As a result of the widespread use of social media for everything, people are increasingly sharing information without understanding if it is phony or genuine. This news has a goal, such as causing a societal problem or undermining an individual's or an organization's dignity. Thus as a result detecting fake news and assessing the quality of news has become quite urgent. Fake news identification aids in determining the authenticity of news pieces that circulate on social media platforms. Hence in our paper, the proposed system's primary goal is to distinguish between true and false news. The methodology that we have devised leans on supervised learning for checking out the legitimacy of news.[8]

II LITERATURE REVIEW

The implementation of false news identification systems has been the subject of numerous investigations.The most popular machine learning algorithms for identifying false news are Support vector machines, Naive Bayes classifiers, Natural Language Processing (NLP) techniques, sentence similarity algorithms, and classification algorithms. Many terms with similar meanings and concepts were used to characterize the credibility of information, including trustworthiness, believability, reliability, accuracy, fairness, and objectivity.[7] But several studies have employed the machine learning method to determine the message's credibility. Using supervised machine learning techniques (Naive Bayes Classification and SVM) and natural language processing, they developed a smart system that turned out to be very innovative for the detection of false news, demonstrating that the suggested framework can detect questionable data with an accuracy of up to 93.5%. Further investigation revealed that they made use of deep neural networks, passive-aggressive classifiers, and naive Bayes classifiers too as examples of machine learning methods. The authors created a database of satirical news (from The Beaverton and The Onion) and actual news (The Toronto Star and The New York Times) in four areas: business, soft skills, civics, and science news), yielding 240 news stories in total. The feature sets that best-represented absurdity, punctuation, and grammar were used for classification and then text data from news stories were transformed into its numeric form using the TF-IDF vectorizer. For the purpose of distinguishing between fake and real news stories, a stylometric (i.e., writing-style) method has been put out. The studies utilized the Buzzfeed dataset2 of partisan and mainstream news pieces, the truthfulness of which was personally annotated. [9]Character and stop word n-grams,

readability indexes, as well as elements like external links and the typical amount of words per paragraph, were among the stylistic features. In one study an interesting work was found which is named “Hoaxy” : A Fraud Online Tracking Platform where the researchers of that work collected data from social media and news websites by using web scraping and web syndication. They measured user activity by counting the user tweets that have been posted, and they determined the URL’s popularity by counting the total number of persons who have posted a tweet. Based on these findings, they drew the conclusion that fact-checking is a more pervasive, grassroots activity while rumors are driven by a small number of active accounts that bear the cost of spreading inaccurate information. All of this work and research has only served to highlight the potential utility of classifiers like Deep Neural Networks, Support Vector Machines, Naive Bayes, Logistic Regression, and Random Forest (DNN) for comparing outcomes, determining accuracy, and enhancing the effectiveness of a classification-based fake news detection tool.[10]

III RESEARCH METHODOLOGY

In this project, we are going to use previous true and false news to train our model so that it can identify any new fake news that comes up.We have compared 5 different classifiers to identify which of the classifier is best for the model.The classifier we used are Logistic Regression, Decision tree classifier, RandomForest classifier, Gradient boosting classifier and Stochastic Gradient Descent classifier.

Figure 1: Flow Chart of our work flow

A Data Collection

Our task was to research form different sources to collect news of all sorts. We collected the data from Kaggle,BuzzFeed, The daily star etc. and made separate files one for true and one for false news.We then combined both the dataset together and then assigned a target column with binary value(1 or 0).1 represents true and 0 represents false, any news containing null values was discarded.We also combined the “title” and body/text of the news into one column and named it “original” which we later used to train and test the model.Our final dataset contained 6 columns which are title, text, subject, date, target and original. Title consists of the headline of our news, Text shows the context of the news, subject determines the main news and date shows the publication of the news. Lastly, our two csv files of both fake and true consists of 25,000 sets of data each.

B Data Analysis

Our collection of data consists of all sorts of news, ranging from world news to government news.Furthermore, analysis of the data collection shows the subject distribution using a graph, the graph showed that out of all the data we had political data was around 70% of it and thus we decided to train our model based on only political news.

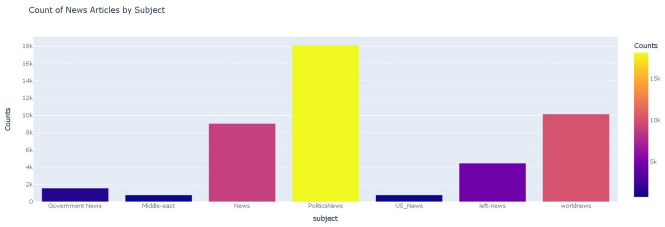

Figure 2: Count of the News Articles by Subjects

We also saw the distribution of true and false data by plotting another graph which shower us that there were more false news than true and thus we had to discard some of the false news so the amount of both false news and true news stays the same and does not provide any skewness to our trained model. Thus, the dataset looks really balanced and hence working on this is pretty easy. Thus we need not work on to make this dataset more balanced, and can safely assume this is a balanced dataset

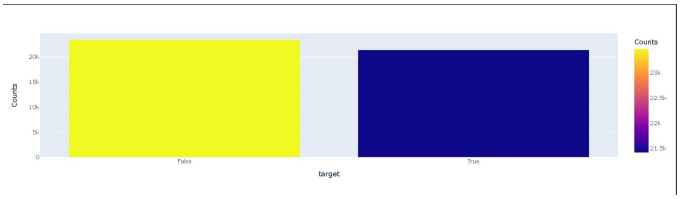

Figure 3: Bar graph of both fake and true news

IV IMPLEMENTATION

A Data Pre-Processing

Before building our model, we had to run through our collected data set to make sure our data is ready for running our algorithms. After researching, we found out our dataset does not contain null values or missing any values. Hence, we can work with all our data. Now our first task comes is to remove any spaces and punctuation marks that are available in our texts. Next step was to import two libraries named NLTK and gensim, which are both open-source

library for unsupervised topic modeling and natural language processing.NLTK is an excellent library for machine-learning based NLP, written in Python by experts from both academia and industry. Python allows you to create rich data applications rapidly, iterating on hypotheses. The combination of Python + NLTK means that you can easily add language-aware data products to your larger analytical workflows and applications. These libraries are useful, as for tasks like text classification, where the text is to be classified into different categories, stopwords are removed or excluded from the given text so that more focus can be given to those words which define the meaning of the text. Stopwords are the most common words in any natural language. For the purpose of analyzing text data and building NLP models, these stopwords might not add much value to the meaning of the document., r6 Generally, the most common words used in a text are “the”, “is”, “in”, “for”, “where”, “when”, “to”, “at” etc. By removing these words we have a created a list of the words which are later combined together to form a phrase, on which we will run our models. Also, news contents which were similar had been labelled to one subject type to avoid ambiguity. After pre-processing our data, we tried to visualize our preprocessed dataset. Here, we are taking our ’original’ column and used WordCLoud for the visual representation. Next, comes a normal

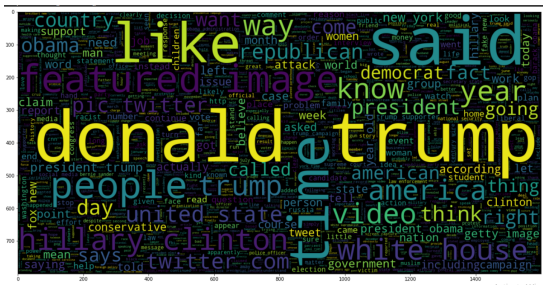

Figure 4: Visual Representation using WordCloud

distribution graph on the total number of words we will use in our data. It has been found average words containing in around 13,753 texts is 100-199. And the maximum word limit was 4582.

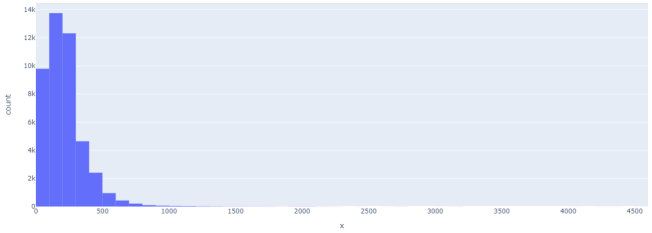

Figure 5: Normal Distribution graph

B Model Building

We have compared 5 different classifiers to identify which of the classifier is best for the model. The classifier we used is Logistic Regression, Decision tree classifier, RandomForest classifier, Gradient boosting classifier, and Stochastic Gradient Descent classifier.

Logistic Regression. When predicting the likelihood of a class based on some dependent variables, the machine learning classification approach known as logistic regression is applied. The logistic regression model, in essence, adds up the input features. Using the Sigmoid function, which transforms numerical data into an expression of probability between 0 and 1.0, it assigns probabilities to discrete outcomes. Two categories are created. After data points have been classified into a class using the Sigmoid function, a hyperplane is utilized as a decision line to divide two groups. The decision boundary can then be used to forecast the kind of upcoming data points.[1] The logistic function is of the form:

$$p(x) = 1/L + e^{-k(x-x0)}$$

Decision Tree Classifier. Both classification and regression tasks need a decision tree. It is organized hierarchically, with a root node, branches, internal nodes, and leaf nodes. A decision tree is created in three steps. Information Gain, Chi-Square, and Gini Impurity. The Gini impurity score ranges from 0 to 1, where 0 indicates that all of the elements belong to one class and 1 indicates that the items are dispersed randomly among the classes.[2] Finding the best nodes that offer the maximum information gain is the crux of information gain. This is computed using a factor known as Entropy. The standardized discrepancies between the observed and expected frequencies of the target variable are represented as the sum of squares in the chi-square statistic.

RandomForest Classifier. Both continuous variables, as in the case of regression, and categorical variables, as in the case of classification, are supported by random forest. In terms of classification issues, it delivers superior outcomes. In Random Forest, n records at random are selected from a data set with k records. [3]Each sample’s decision tree is built separately. The output of each decision tree will be produced. For classification and regression, the final result is based on the majority vote or average, respectively.

Gradient Boosting Classifier. When we wish to reduce bias error, we typically utilize the gradient boosting algorithm. It is a method of merging different weak predictors, usually Decision Trees, to create an additive predictive model. We begin with a leaf, which stands for each person’s initial prediction. Next, we employ the Logistic Function to transform this into a probability.[4] Then, for each observation, we shall compute the residuals. Building a Decision Tree to forecast the residuals is the next stage.

Stochastic Gradient Descent Classifierr. The main objective of the stochastic gradient descent classifier is to locate the local minima of any differential function. setting up the random w words. The predictions will be calculated using the w term algorithm. Determine the mean square error between predicted values and actual values.[5] utilizing the previous value of the parameter and the mean square error, determine the updated value of the parameter (w). Until convergence, keep calculating the parameter’s updated value and forecast.

V RESULTS AND ANALYSIS

The results we got can be separated into two segments. The first part involves training the model with specific variables from the dataset and checking its accuracy score. Our dataset contains columns named 'Title'and 'Text', and these columns were separated and used individually to train the model. A third test was carried out by combining both columns to see if it yields a better result. Finally, we compared the accuracies found from these three different approaches to determine the best method.

Firstly, the 'Title' column was taken as the X-train, and Logistic Regression algorithm was run on it. An accuracy score of 94.75% was recorded. The confusion matrix of the result was also generated to show the details of the outcome. The true-positive, true-negative, false-positive, and false-negative values displayed in the matrix were used to analyze the model's performance.

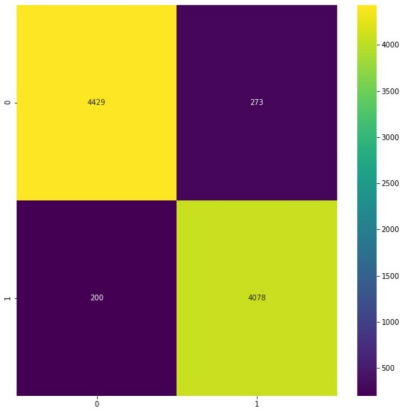

Figure 6: Confusion matrix generated when using the 'Title' column

For the second approach, we took the 'Text' column as our input to test and train the model. To keep the experiment fair, we made sure to keep all the variables used in the Logistic Regression algorithm exactly the same for all three approaches. The accuracy in this experiment came to be 99.53

For the final part of the experiment, we combined both the 'Title' and 'Text' columns and checked how the model performs with the joined dataset. Like all previous tests, Logistic Regression was implemented, followed by a confusion matrix. The accuracy found was 99.67%, which is the highest of the three approaches. Similarly, the confusion matrix also mimicked the same outcome with only 28 data from the entire dataset being falsely predicted.

Next comes to the second segment of our results analysis. In addition to Logistic Regression, we also tried out a few other machine learning algorithms and compared their accuracy values. The algorithms used were Decision Tree, Gradient Boosting, Random Forest, and Stochastic Gradient Descent. The table below shows the ML algorithms used with their respective accuracy scores.

The information in the table above shows that Logistic Regression predicts fake news with the highest accuracy. Thus, we can conclude that using Logistic Regression, for our model, is the best option to get highly accurate results.

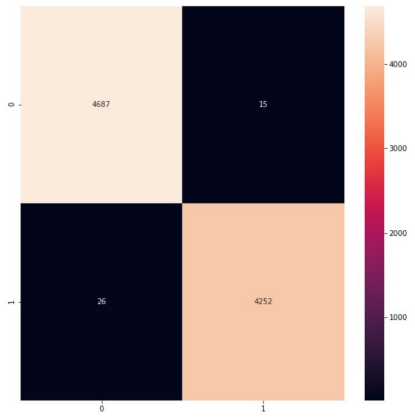

Figure 7: Confusion matrix generated when using the 'Text' column

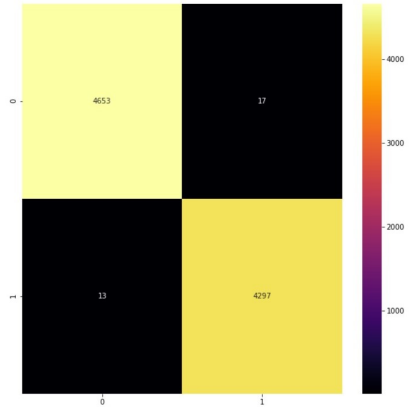

Figure 8: Confusion matrix generated when using both the 'Title' and 'Text' columns

Model	Accuracy
Logistic Regression	99.67%
Decision Tree	90.79%
Gradient Boosting	82.16%
Random Forest	93.51%
Stochastic Gradient Descent	94.62%

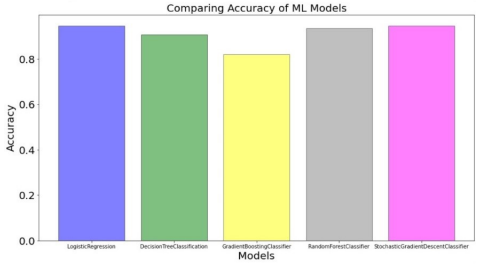

Figure 9: Accuracies of all Machine Learning Algorithms Tested

VI CONCLUSION

A Challenges

One of the challenges which we have faced is the availability of a wide range of topics or genres of news in the papers such as sports, business or entertainment. Usually, the most important topics found on online news pages or articles are based on government, politics, and health thus we are missing out on other major topics. The other most difficult challenge is not having a dynamic dataset. The issue here arises when new data is not constantly fed into the dataset by a reliable news media source. So if every new data is checked through our model it will come up as fake news even if it is genuine since model training has not been done with the updated dataset.

B Future Work

A greater accuracy than the current one can be achieved if we use the author and the time of publication along with text and headline for training our model. Also in the future, we should try to collect other genres of news as well such as entertainment, business and etc. Ultimately, we can also consider including other languages.

C Epilogue

We should investigate the truth behind fake news before believing it and spreading it via social media. In our work, we use both text and Headline of news and run them through Logistic Regression for predicting whether the news is fake or accurate. According to the study's findings, the system can detect whether the news is fake or real with an accuracy of up to 99.67%. Our research establishes a new framework for determining the veracity of the news.

ACKNOWLEDGMENTS

This work was supported by Miss Sifat E. Jahan, respected faculty of BRAC University. Additionally, we want to thank Department of CSE for helping to conduct our research. We also thank for contributing to the news resources which helped in data collection.

REFERENCES

[1] How does logistic regression work? KDnuggets. (n.d.). Retrieved September 6, 2022, from https://www.kdnuggets.com/2022/07/logistic-regression-work.html

[2] Kurama, V. (2021, April 9). A complete guide to decision trees. Paperspace Blog. Retrieved September 6, 2022, from https://blog.paperspace.com/decision-trees/

[3] Random Forest: Introduction to random forest algorithm. Analytics Vidhya. (2022, June 21). Retrieved September 6, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[4] Gradient boosting trees for classification: A beginner's guide. Affine. (2022, April 18). Retrieved September 6, 2022, from https://affine.ai/gradient-boosting-trees-for-classification-a-beginners-guide/

[5] Verma, Y. (2022, March 11). A beginner's guide to stochastic gradient descent from scratch. Analytics India Magazine. Retrieved September 6, 2022, from https://analyticsindiamag.com/a-beginners-guide-to-stochastic-gradient-descent-from-scratch/

[6] How to remove Stopwords in python: Stemming and lemmatization. Analytics Vidhya. (2020, December 23). Retrieved September 6, 2022, from https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/

[7] Fake news detection and prediction using machine learning algorithms. (n.d.). Retrieved September 6, 2022, from https://www.researchgate.net/publication/353824797Fake News Detection and Prediction Using Machine Learning Algorithms

[8] Supervised learning for fake news detection - sentic.net. (n.d.). Retrieved September 6, 2022, from https://sentic.net/supervised-learning-for-fake-news-detection.pdf

[9] (PDF) automatic detection of fake news - researchgate. (n.d.). Retrieved September 6, 2022, from https://www.researchgate.net/publication/319255985 Automatic Detection of Fake News

[10] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020, October 17). Fake news detection using machine learning ensemble methods. Complexity. Retrieved September 6, 2022, from https://www.hindawi.com/journals/complexity/2020/8885861/