FAKE NEWS DETECTION USING SUPERVISED LEARNING METHOD

A Project-II Report

Submitted in partial fulfillment of requirement of the

Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

BY

Swapnesh Jain
EN16CS301272

Under the Guidance of

Prof. (Dr.) Ruchi Patel
Prof. Sachin Solanki

Department of Computer Science & Engineering
Faculty of Engineering
MEDI-CAPS UNIVERSITY, INDORE- 453331

May, 2020

FAKE NEWS DETECTION USING SUPERVISED LEARNING METHOD

A Project-II Report

Submitted in partial fulfillment of requirement of the

Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

BY

Swapnesh Jain
EN16CS301272

Under the Guidance of
Prof. (Dr.) Ruchi Patel
Prof. Sachin Solanki

Department of Computer Science & Engineering
Faculty of Engineering
MEDI-CAPS UNIVERSITY, INDORE- 453331

May, 2020

Report Approval

The project work **"Fake News Detection Using Supervised Learning Method"** is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as prerequisite for the Degree for which it has been submitted.

It is to be understood that by this approval the undersigned do not endorse or approved any statement made, opinion expressed, or conclusion drawn there in; but approve the "Project Report" only for the purpose for which it has been submitted.

Internal Examiner

Name:

Designation

Affiliation

External Examiner

Name:

Designation

Affiliation

Declaration

I hereby declare that the project entitled **"Fake News Detection Using Supervised Learning Method"** submitted in partial fulfillment for the award of the degree of Bachelor of Technology in 'Computer Science & Engineering' department completed under the supervision of **Prof. Sachin Solanki, Department of Computer Science & Engineering,** Faculty of Engineering, Medi-Caps University Indore and **Prof. (Dr.) Ruchi Patel, Department of Computer Science & Engineering,** Faculty of Engineering, Medi-Caps University Indore is an authentic work.

Further, I declare that the content of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for the award of any degree.

Swapnesh Jain

May 13, 2020

Certificate

We, **Prof. Sachin Solanki and Prof. (Dr.) Ruchi Patel** certify that the project entitled **"Fake News Detection Using Supervised Learning Method"** submitted in partial fulfillment for the award of the degree of Bachelor of Technology by **Swapnesh Jain (EN16CS301272)** is the record carried out by him under our guidance and that the work has not formed the basis of award of any other degree elsewhere.

Prof. Sachin Solanki

Department of Computer Science & Engineering

Medi-Caps University, Indore

Prof. (Dr.) Ruchi Patel

Department of Computer Science & Engineering

Medi-Caps University, Indore

Dr. Suresh Jain

Head of the Department

Department of Computer Science & Engineering

Medi-Caps University, Indore

Acknowledgements

I would like to express my deepest gratitude to Honorable Chancellor, **Shri R C Mittal,** who has provided me with every facility to successfully carry out this project, and my profound indebtedness to **Prof. (Dr.) Sunil K Somani,** Vice Chancellor, Medi-Caps University, whose unfailing support and enthusiasm has always boosted up my morale. I also thank **Prof. (Dr.) D K Panda,** Dean, Faculty of Engineering, Medi-Caps University, for giving me a chance to work on this project. I would also like to thank my Head of the Department, **Dr. Suresh Jain** for his continuous encouragement for betterment of the project.

I would also like to express my heartfelt gratitude to my Internal Guide, **Prof. (Dr.) Ruchi Patel and Prof. Sachin Solanki,** Department of Computer Science & Engineering, Medi-Caps University, without whose continuous help and support, this project would ever have reached to the completion.

It is their help and support, due to which I became able to complete the design and technical report.

Without their support this report would not have been possible.

Swapnesh Jain (EN16CS301272)
B.Tech. IV Year
Department of Computer Science & Engineering
Faculty of Engineering
Medi-Caps University, Indore

Abstract

With the advent of social media, information has never been more accessible than today. But as much as it is accessible, there is a flood of information at the time where virtually everyone is a content creator. In this age where data is the new oil, the amount of data produced every day on social media has caused certain troubles too. One of the most notorious trouble thus created is that of “fake news”. While social media has proven to be a voice of the previously voiceless and has enabled those to dismantle old one-sided narratives but on the other hand, it has also enabled people to monger fear, incite violence and spread misinformation. The most recent effect of fake news can be seen in the violence happening in New Delhi, India, where a systematic spread of misinformation led to riots against the majority community of India.

In this paper, a comparative study has been made for the efficiency of various models in correctly identifying fake news along with their true positives and true negatives. The aim here is to apply count vector and tf-idf vector to four different machine learning methods such as Naïve Bayes, Logistic Regression, Random Forest and XGBoost on two different datasets namely Kaggle and Liar. Based on the results obtained XGBoost along with count vector gave the highest accuracy in predicting fake news.

Keywords: Count Vector, Fake News, Logistic Regression, Naïve Bayes, Random Forest, Tf-Idf Vector, True Positives, True Negatives, XGBoost.

Table of Contents

List of Figures

List of Tables

Chapter-1

Introduction

1.1 Introduction

With the information revolution now being a real thing and virtually every person with a smartphone in hand and a working internet connection being a content creator everyone has to deal with a new menace of fake news almost every day with social media being its breeding ground. With more and more people being introduced to social media every day and it is a great influence on the opinion-forming process of the people, there are people with vested agendas who want to use these information as a means to propel their agenda.

This has led to the creation of many websites that publish articles containing half-truths or full lies. Some websites publish fake news almost exclusively to push propaganda (mostly political) to influence people in a certain way. The fake news industry is a global issue as well as a challenge for the world with many established media houses also losing credibility over the years after being repeatedly caught in fake news.

Many scientists believe that this problem can be dealt by using Machine Learning techniques with Artificial Intelligence. Hence this paper describes a comparative study of four most popular machine learning methods to identify and determine which method produces the best results while detecting fake news.

1.2 Literature Review

Fake news is a major challenge. Most challenging part of fake news detection is the detection of deceptive languages which is done by statistical methods. This issue becomes even more serious while dealing with reviews obtained from interviews on television, social media posts like those on Facebook and Twitter. Maniz Shrestha [1] in his work combined sentiment analysis with network metadata to detect fake news. He trained Random Forest classifier which gave him the f1-score of more than 88%. He also designed a scrapping tool to gather news related articles from different sources.

Kyeong-Hwan Kim and Chang-sung Jeong [2] of Korea University co-authored a paper on article abstraction in which they created a factual database by collecting obvious facts

of human's decisions. Their system search for the articles related to the news in their factual database to verify whether the news is reliable or not.

Rohit Kumar Kaliyar [3] implemented deep neural network for detecting fake news which include different machine learning models along with different deep learning models which evaluate their performance in identifying fake news. Recent work by M. S. Mokhtar et al [4] include an integrated web service model that accepts news input or URL from the user which then checks for the truth level of the news. Also Mykhailo Granik and Volodymyr Mesyura [5] in their study have focused on the detection of fake news by training the Naïve Bayes classifier which was then tested on Facebook posts. Their model achieved classification accuracy of 74% on the test data. In other studies work like evaluation of different machine learning methods using tf-idf and probabilistic context-free grammar [6], fake news detection on social media networks to filter out sites with false and misleading news [7], and hybrid text classification [8] to deal with the classification of fake news were carried out.

1.3 Objectives

The main objective to carry out this project is:

1. To design an algorithm and model that will detect whether a news is fake or not.
2. To evaluate the performance of different machine learning models using confusion matrix and accuracy.
3. Comparative study between the accuracies and confusion matrices of different models for different datasets.

1.4 Problem Domain

One of the problem is that most of the news are gathered from various sources present on the web, therefore the information obtained from these news cannot be relied on and these sources cannot be trusted because the true origin of most of these news (present on the Internet) is not known.

Also over the past few years spreading of rumours and misinformation has reached a point that it has begun to affect social issues and political problems.

Another problem is that the amount of time spent on social media has increased at an alarming rate. Thus most of the fake news are acquired from these sources. Social media provides anonymity while expressing out opinions which greatly reduces the authenticity of news received from these sources as compared to a newspaper or any other trusted media.

1.5 Solution

The solution for all these problems lies in designing and implementing various machine learning models and training them using dataset (containing news articles that are already classified into fake and real ones), that can be used to predict whether a given news is fake or real. And the probability for the news to be true can be checked using the accuracy obtained from these predicted models.

Chapter-2

Methodologies Used

For classification problem many different supervised learning techniques are present such as Naïve Bayes, Decision Trees, Support Vector Machine, Gradient Descent, K-Nearest Neighbors, K-Fold cross validation, Neural Networks, etc. Some of these popular techniques have been used in this study which are as follows:

2.1 Naïve Bayes (NB)

Naive Bayes algorithm is a simple but an efficient technique for the construction of classifiers. It consists of models that label problem instances as classes, which are represented as feature-vectors. There is not just a single algorithm to train all such classifiers, but there exists a family of algorithms which all are based on a common principle. All NB classifiers work on the assumption that for a given class variable the value of a feature does not depend upon the value of any other feature. Despite their naive and primitive design and seemingly oversimplified assumptions, NB classifiers have been known to function quite well in many complex scenarios that exist in the real-world. An advantage of NB classifier is that only a small number of training data is required to make an estimation of the parameters that would be necessary for classification.

2.2 Logistic Regression (LR)

Logistic Regression algorithm is applied when the dependent or the target variable is categorical in nature. In its basic form LR is a statistical model which makes uses of a logistic function for modelling a binary dependent variable. Models based on analogues technique that use a different kind of sigmoid function instead of using the logistic function can also be considered for logistic regression, one such example is the probity model. The characteristic feature that defines the logistic model is that when value of one of the independent variable increases multiplicatively, the probability of the second variable gets scaled at a constant rate, where each independent variable has its own parameter; this feature generalizes the odds ratio for a binary dependent variable also.

2.3 Random Forest (RF)

As the name implies, Random forest is a model that consists of numerous individual and independent decision trees that function simultaneously as an ensemble. Each individual trees

in the RF returns a class as a prediction and the class that attains the highest number of votes becomes the prediction of the entire model. The distinguishing feature of RF is that when a fairly large number of independent trees work together, they outperform any other standalone model. The trees also enforce canopy effect like that of a real forest and protect each other from their respective individual errors [9].

2.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (herein after referred to as XGBoost) is an ensemble machine learning algorithm that is based on the concept of decision trees and works on the framework of gradient boosting. In such prediction problems that deal with the data in the unstructured form (images, text, etc.) artificial neural networks tend to outdo the performance of all other algorithms and frameworks working with similar problems. However, while dealing with small-to-medium scale structured or tabular data, algorithm based on decision tree concept are considered to be the most suited model and are widely recommended for this purpose [10]. XGBoost is one such gradient boosting library that is highly optimized and distributed and is designed to be highly efficient, flexible and portable.

Chapter-3

Report on Present Investigation

3.1 Dataset

To implement each method two different datasets are gathered for this projects, namely Kaggle and Liar.

The first dataset is of Kaggle which includes 16,548 human labeled news articles in the train set described by five columns namely ‘id’, ‘title’, ‘author’, ‘text’ and ‘label’ (categorized as 0, indicating real news and 1, indicating fake news). The test dataset of Kaggle consists of 4,137 news articles described by four columns namely ‘id’, ‘title’, ‘author’ and ‘text’. This test set contains news without label. And the submit dataset of Kaggle consists of same 4,137 news articles, as of test dataset, described by two columns namely ‘id’ and ‘label’ but with label to compute the accuracy of each method.

Second dataset is of Liar which includes 10,240 human labeled news articles in the train set described by two columns namely ‘statement’ and ‘label’ (categorized as true and false). The test dataset of Liar consists of 2,551 human labeled news articles described by two columns namely ‘statement’ and ‘label’ to compute the accuracy of each model.

3.2 Platform Specification

Anaconda platform has been used as a python development environment to build and train machine learning models for detecting fake news using built-in packages and libraries.
The project has been carried out using Python language because it provides a great choice of libraries and packages to carry out most of the tasks to build machine learning models. Also Python has a great community support because most of the programmers contribute to help each other.

3.3 Pre-processing

First both the datasets are observed to look for what type of data each methods are dealing with. But upon closely observing datasets it was found that both the Kaggle and Liar datasets contain missing values and incomplete news articles. But Kaggle dataset along with missing values and incomplete news articles contains news from different languages also. So to increase the efficiency of each models these missing values along with the news articles

which are incomplete and of different languages are eliminated. The text field plays an important role in labelling and therefore text field containing blank columns have also been deleted.

After eliminating these type of data all the special characters (or say punctuation marks and tags) along with the stop-words are removed. The stop-words are those words that do not contribute much in predicting whether the news is real or fake, instead they just introduce confusion to each model. Some of these commonly used stop-words are "a", "the", "of", "I", "you", "it", "and", etc. Rather machine would not want these words taking up space in the database, or taking up valuable processing time. Therefore eliminating them will improve the efficiency of the model. After the soft-words are removed, stemming was carried out to reduce the words to their roots.

3.4 Proposed Solution

This section consists of important steps (see Fig. 3.1) for detecting fake news using different machine learning methods, which are as follows:

Step 1: Two different datasets, Kaggle and Liar, are gathered to implement each method.

Step 2: Then cleaning techniques such as Tokenization, Stemming and removing punctuation marks, tags, and stop-words are used on each dataset to clean them.

Step 3: After cleaning, two different pre-processing techniques such as CountVectorizer (count of terms in vector) and term frequency-inverse document frequency (tf-idf) are used on each dataset to build the vocabulary of the count vectors.

Step 4: The methods then are trained using Naive Bayes, Logistic regression, Random Forest and XGBoost.

Step 5: Then performance was evaluated for these different machine learning methods and compared using their accuracy and confusion matrix.

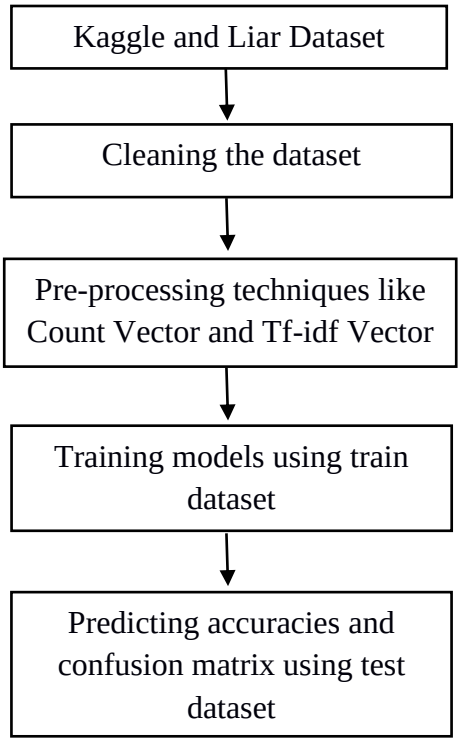

Fig. 3.1. Flowchart of Proposed Solution

In the above proposed solution, steps 2 and 3 are important steps because these steps are not only normalizing the dataset but also converting the textual news data into numerical vectors as these machine learning methods can be trained using numerical data only. In the step 2 the dataset is first cleaned because when the dataset is gathered from different sources they are in raw format which is not acceptable by the methods. In the step 3 CountVectorizer is used to count the frequency of each word occurred in the news. TF-IDF is also used for information retrieval to determine the importance of individual terms in the set of text documents. TF-IDF's value increases proportionately with the increase in occurrence of the given word in a given document but often falls down with the frequency of the words in the corpus.

TF-IDF can be computed as the product of term frequency and inverse document frequency. Term frequency can be computed as the ratio of number of times a term appears in a document to the total number of terms in the document and Inverse Document Frequency can be computed as the log of ratio of number of documents to the number of documents that contain the word.

Once the textual data is converted into the numerical vectors (using CountVectorizer and TF-IDF) the methods are trained using these vectors so that the output of the test dataset can be predicted to compare the accuracies between these four different methods to determine their efficiency. Also a confusion matrix is generated which classifies these news as positive or negative to evaluate their performance.

In the confusion matrix shown in Table 3.1, rows represent the number of classifications predicted by the model, while the columns represent the number of actual classifications in the test data.

Table 3.1. Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

Where, TP is defined as actual real news which are correctly predicted as real,

FP is defined as actual false news which is incorrectly predicted as real,

FN is defined as actual real news which is incorrectly predicted as false, and

TN is defined as actual false news which is correctly predicted as false.

Accuracy is the ratio of sum of predictions that are correctly classified by the model to the total number of samples. Accuracy can be computed using the equation (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

3.5 Results and Discussions

The model was trained and tested for both the datasets. In this project two techniques for pre-processing, namely ‘CountVectorizer’ and ‘TF-IDF’, and four methods of machine learning (discussed in Methodologies) are used and confusion matrix along with accuracies are obtained and tabulated.

3.5.1 Experimental Result on Kaggle Dataset

Table 3.2 describes in detail the labels obtained from various prediction methods and also the extent of correctness. CountVectorizer along with various prediction methods was applied on Kaggle dataset to obtain this table.

With Naïve Bayes classifier, 3489 correct and 368 incorrect predictions are obtained. With Logistic Regression, 3680 correct and 177 incorrect predictions are obtained. On applying Random Forest, 3529 correct and 328 incorrect predictions are obtained. Finally while applying XGBoost, 3716 correct and 141 incorrect predictions are obtained.

Table 3.2. Confusion Matrix when CountVectorizer was used as pre-processing technique on Kaggle dataset

Classification Algorithm	Label	Actual Positive	Actual Negative
Naïve Bayes	Predicted Positive	1938	98
	Predicted Negative	270	1551
Logistic Regression	Predicted Positive	1957	79
	Predicted Negative	98	1723
Random Forest	Predicted Positive	1975	61
	Predicted Negative	267	1554
XGBoost	Predicted Positive	1961	75
	Predicted Negative	66	1755

Table 3.3 contains the consolidated data for accuracy of various prediction methods with CountVectorizer on Kaggle dataset in percentage point terms.

Model obtained 90.45% accuracy using Naïve Bayes, 95.41% accuracy using Logistic regression, 91.49% accuracy using Random Forest and 96.34% accuracy using XGBoost.

Table 3.3. Accuracy table for CountVectorizer technique on Kaggle dataset

Classification Algorithms	Accuracy %
Naïve Bayes	90.45
Logistic Regression	95.41
Random Forest	91.49
XGBoost	**96.34**

Table 3.4 describes in detail the labels obtained from various prediction methods and also the extent of correctness. TF-IDF Vectorizer along with various prediction methods was applied on Kaggle dataset to obtain this table.

With Naïve Bayes classifier, 3326 correct and 531 incorrect predictions are obtained. With Logistic Regression, 3668 correct and 189 incorrect predictions are obtained. On applying Random Forest, 3557 correct and 300 incorrect predictions are obtained. Finally while applying XGBoost, 3687 correct and 170 incorrect predictions are obtained.

Table 3.4. Confusion Matrix when TF-IDF was used as pre-processing technique on Kaggle dataset

Classification Algorithm	Label	Actual Positive	Actual Negative
Naïve Bayes	Predicted Positive	2016	20
	Predicted Negative	511	1310
Logistic Regression	Predicted Positive	1954	82
	Predicted Negative	107	1714
Random Forest	Predicted Positive	1962	74
	Predicted Negative	226	1595
XGBoost	Predicted Positive	1955	81
	Predicted Negative	89	1732

Table 3.5 contains the consolidated data for accuracy of various prediction methods with TF-IDF Vectorizer on Kaggle dataset in percentage point terms.

Model obtained 86.23% accuracy using Naïve Bayes, 95.09% accuracy using Logistic regression, 92.22% accuracy using Random Forest and 95.59% accuracy using XGBoost.

Table 3.5. Accuracy table for TF-IDF technique on Kaggle dataset

Classification Algorithms	Accuracy %
Naïve Bayes	86.23
Logistic Regression	95.09
Random Forest	92.22
XGBoost	**95.59**

3.5.2 Experimental Result on Liar Dataset

Table 3.6 describes in detail the labels obtained from various prediction methods and also the extent of correctness. CountVectorizer along with various prediction methods was applied on Liar dataset to obtain this table.

With Naïve Bayes classifier, 1567 correct and 984 incorrect predictions are obtained. With Logistic Regression, 1525 correct and 1026 incorrect predictions are obtained. On applying Random Forest, 1573 correct and 978 incorrect predictions are obtained. Finally while applying XGBoost, 1570 correct and 981 incorrect predictions are obtained.

Table 3.6. Confusion Matrix when CountVectorizer was used as pre-processing technique on Liar dataset

Classification Algorithm	Label	Actual Positive	Actual Negative
Naïve Bayes	Predicted Positive	586	583
	Predicted Negative	401	981
Logistic Regression	Predicted Positive	590	579
	Predicted Negative	447	935
Random Forest	Predicted Positive	561	608
	Predicted Negative	370	1012
XGBoost	Predicted Positive	482	687
	Predicted Negative	294	1088

Table 3.7 contains the consolidated data for accuracy of various prediction methods with CountVectorizer on Liar dataset in percentage point terms.

Model obtained 61.42% accuracy using Naïve Bayes, 59.78% accuracy using Logistic regression, 61.66% accuracy using Random Forest and 61.54% accuracy using XGBoost.

Table 3.7. Accuracy table for CountVectorizer technique on Liar dataset

Classification Algorithms	Accuracy %
Naïve Bayes	61.42
Logistic Regression	59.78
Random Forest	**61.66**
XGBoost	61.54

Table 3.8 describes in detail the labels obtained from various prediction methods and also the extent of correctness. TF-IDF Vectorizer along with various prediction methods was applied on Liar dataset to obtain this table.

With Naïve Bayes classifier, 1528 correct and 1023 incorrect predictions are obtained. With Logistic Regression, 1561 correct and 990 incorrect predictions are obtained. On applying Random Forest, 1573 correct and 978 incorrect predictions are obtained. Finally while applying XGBoost, 1535 correct and 1016 incorrect predictions are obtained.

Table 3.8. Confusion Matrix when TF-IDF was used as pre-processing technique on Liar dataset

Classification Algorithm	Label	Actual Positive	Actual Negative
Naïve Bayes	Predicted Positive	366	803
	Predicted Negative	220	1162
Logistic Regression	Predicted Positive	531	638
	Predicted Negative	352	1030
Random Forest	Predicted Positive	515	654
	Predicted Negative	324	1058
XGBoost	Predicted Positive	479	690
	Predicted Negative	326	1056

Table 3.9 contains the consolidated data for accuracy of various prediction methods with TF-IDF Vectorizer on LIAR dataset in percentage point terms.

Model obtained 59.89% accuracy using Naïve Bayes, 61.19% accuracy using Logistic regression, 61.66% accuracy using Random Forest and 60.17% accuracy using XGBoost.

Table 3.9. Accuracy table for TF-IDF technique on Liar dataset

Classification Algorithms	Accuracy %
Naïve Bayes	59.89
Logistic Regression	61.19
Random Forest	**61.66**
XGBoost	60.17

This study has used two pre-processing algorithms and four prediction methods in all permutations on two datasets namely Kaggle and Liar. While testing the permutation on Kaggle dataset it was found that the highest accuracy was achieved when CountVectorizer technique was applied and XGBoost prediction method was used. The accuracy achieved by this model was 96.34%.

The same set of permutation was used to predict fake news on Liar dataset. The highest accuracy was achieved when CountVectorizer technique was applied and Random forest prediction method was used. The accuracy achieved by this model was 61.66%.

Chapter-4

Conclusion and Future Scope

With the proliferation of social media, more and more people are gaining access to news from unconventional sources like social media rather than traditional main stream media. This has led to mushrooming of fake news platforms all over the world.

Therefore this paper analyses various text pre-processing techniques and classification methods that can predict whether a news is fake or not. In this experiment it was found that XGBoost classifier with CountVectorizer technique achieved the highest of accuracy of 96.34%, therefore it can be used to predict whether a news is fake or real.

The project has also exposed me to the latest technology area in the field of machine learning. This project deals with the analysis of the news articles to predict whether a news article is fake or real, to solve the problems being faced while dealing with the truth of news and information.

In the future this model can be extended to sentiment based model and also algorithms like Recursive Neural Networks (RNN) can be applied to further improve the efficiency. This project can be expanded further to include fact-checking and deep syntax analysis, as well as recommending similar credible articles. Convolutional Neural Network (CNN) can be used on image type of news to detect whether they are real news or misleading news. Also further features, such as semantic features, parts-of-speech tagging, etc can be expanded to increase the accuracy.

Literature Citation

1. Maniz Shrestha, 2018. "Detecting Fake News with Sentiment Analysis and Network Metadata," Earlham college, Richmond.
2. K. Kim, C. Jeong, 2019. "Fake News Detection System using Article Abstraction," 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chonburi, Thailand, p. 209-212.
3. Rohit K. Kaliyar, 2018. "Fake News Detection Using A Deep Neural Network," 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, p. 1-7.
4. M. S. Mokhtar, Y. Y. Jusoh, N. Admodisastro, N. C. Pa, A. Y. Amruddin, October 2019. "Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning," International Journal of Engineering and Advanced Technology (IJEAT), Volume-9, Issue-1, p. 2407-2410.
5. M. Granik, V. Mesyura, 2017. "Fake news detection using naive Bayes classifier," IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, p. 900-903.
6. Shlok Gilda, 2017. "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," IEEE 15th Student Conference on Research and Development (SCOReD), Putrajaya, p. 110-115.
7. M. Aldwairi, A. Alwahedi, 2018. "Detecting Fake News in Social Media Networks," The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018), Volume 141, Procedia Computer Science, Abu Dhabi, p. 215-222.
8. P. Kaur, R. S. Boparai, D. Singh, June 2019. "Hybrid Text Classification Method for Fake News Detection," International Journal of Engineering and Advanced Technology (IJEAT), Volume-8, Issue-5, p. 2388-2392.
9. "Understanding Random Forest", Medium, 2020. Accessed on: 20-March-2020.
10. V. Morde, "XGBoost Algorithm: Long May She Reign!" Medium, 2020. Accessed on: 20-March-2020.

Bibliography and References

- https://portfolios.cs.earlham.edu/wp-content/uploads/2018/12/Fake_News_Capstone.pdf
- https://ieeexplore.ieee.org/document/8864154
- https://ieeexplore.ieee.org/document/8777343
- https://www.ijeat.org/wp-content/uploads/papers/v9i1/A2633109119.pdf
- https://ieeexplore.ieee.org/document/8100379
- https://ieeexplore.ieee.org/document/8305411
- https://www.sciencedirect.com/science/article/pii/S1877050918318210
- https://www.ijeat.org/wp-content/uploads/papers/v8i5/E7622068519.pdf
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d