FAKE NEWS DETECTION USING LOGISTIC REGRESSION & MULTINOMIAL NAIVE BAYES

Abhishek Singh[1], Aditya Ugale[2], Niraj Shah[3], Prof. Amruta Sankhe[4]

[1,2,3]Student, Information Technology Department, Atharva college of engineering, Maharashtra, India
[4]Asst. Professor, Atharva College of Engineering, Maharashtra, India

Abstract – *In today's worlds where people are more reliable on the news which are available online as it's convenient for them. As the use of the internet is increasing so thus the spread of fake news also. As the spread of such fake news can be intentional or unintentional but this affects society. Thus, an increasing number of fake news has to be controlled by using the computational tool which predicts such misleading information as if it is fake or real. In this article, we have focused on developing such computational tool to help classify news using two different algorithms. Which helps the model to be more trustworthy We will describe the pre-processing, feature extraction, classification and prediction process in detail. We've used Logistic Regression and Multinomial language processing techniques to classify fake news. The pre-processing functions perform some operations like tokenizing, lemmatization and exploratory data analysis like response variable distribution and data quality check (i.e., null or missing values). Simple Count Vectorization, TF-IDF is used as feature extraction techniques. The logistic regression and Multinomial model are used as a classifier for fake news detection with a probability of truth.*

Key Words**: Fake news detection, Logistic regression, TF-IDF, count vectorization, Multinomial Naïve Bayes, NLP, feature selection.**

1. INTRODUCTION

These days' fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is "fake news" but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints.

Fortunately, there are a number of computational techniques that can be used to mark certain articles as fake on the basis of their textual content. Majority of these techniques use fact checking websites such as "PolitiFact" and "Snopes." There are a number of repositories maintained by researchers that contain lists of websites that are identified as ambiguous and fake. However, the problem with these resources is that human expertise is required to identify articles/websites as fake.

As human beings, when we read a sentence or a paragraph, we can interpret the words with the whole document and understand the context. In this project, we teach a system how to read and understand the differences between real news and the fake news using concepts like natural language processing, NLP and machine learning and prediction classifiers like the Logistic regression and multinomial Naïve bayes which will predict the truthfulness or fake-news of an article. We have also made a sentimental analysis of the news or article as to get as its positive or negative news.

2. LITERATURE REVIEWS

In general, Fake news could be categorized into three groups. The first group is fake news, which is news that is completely fake and is made up by the writers of the articles. The second group is fake satire news, which is fake news whose main purpose is to provide humour to the readers. The third group is poorly written news articles, which have some degree of real news, but they are not entirely accurate. In short, it is news that uses, for example, quotes from political figures to report a fully fake story. Usually, this kind of news is designed to promote certain agenda or biased opinion.

"Fake News Detection" Akshay Jain, Amey Kasbe[1], 2018 IEEE. Information preciseness on Internet, especially on social media, is an increasingly important concern, but web-scale data hampers, ability to identify, evaluate and correct such data, or so called "fake news," present in these platforms. In this paper, we propose a method for "fake news" detection and ways to apply it on Facebook, one of the most popular online social media platforms. This method uses Naive Bayes classification model to predict whether a post on Facebook will be labelled as REAL or FAKE. The results may be improved by applying several techniques that are discussed in the paper. Received results suggest, that fake news detection problem can be addressed with machine learning methods.

Fake news detection in social media Kelly Stahl, 2018 California State University Stanislaus[2]. Due to the exponential growth of information online, it is becoming impossible to decipher the true from the false. Thus, this leads to the problem of fake news. This research considers previous and current methods for fake news detection in textual formats while detailing how and why fake news exists in the first place. This paper includes a discussion on Linguistic Cue and Network Analysis approaches, and proposes a three-part method using Naïve Bayes Classifier, Support Vector Machines, and Semantic Analysis as an accurate way to detect fake news on social media.

Survey on Automated System for Fake News Detection using NLP & Machine Learning Approach, Subhadra Gurav, Swati Sase, Supriya Shinde, Prachi Wabale, Sumit Hirve[3] The large use of social media has tremendous impact on our society, culture, business with potentially positive and negative effects. Now-a-days, due to the increase in use of online social networks, the fake news for various commercial and political purposes has been emerging in large numbers and widely spread in the online world. The existing systems are not efficient in giving a precise statistical rating for any given news. Also, the restrictions on input and category of news make it less varied. This paper develops a method for automating fake news detection for various events. We are building a classifier that can predict whether a piece of news is fake based on data sources, thereby approaching the problem from a purely NLP perspective.

Fake News Detection Using Machine Learning approaches: A systematic Review, Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita[4] The easy access and exponential growth of the information available on social media networks has made it intricate to distinguish between false and true information. The easy dissemination of information by way of sharing has added to exponential growth of its falsification. The credibility of social media networks is also at stake where the spreading of fake information is prevalent. Thus, it has become a research challenge to automatically check the information viz a viz its source, content and publisher for categorizing it as false or true. Machine learning has played a vital role in classification of the information although with some limitations. This paper reviews various Machine learning approaches in detection of fake and fabricated news. The limitation of such and approaches and improvisation by way of implementing deep learning is also reviewed.

Detecting Fake News using Machine Learning and Deep Learning Algorithms. Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq[5]. Social media interaction especially the news spreading around the network is a great source of information nowadays. From one's perspective, its negligible exertion, straightforward access, and quick dispersing of information that led people to look out and eat up news from internet-based life. Twitter being a standout amongst the most well-known ongoing news sources additionally ends up a standout amongst the most dominant news radiating mediums. It is known to cause extensive harm by spreading bits of gossip previously. Online clients are normally vulnerable and will, in general, perceive all that they run over web-based networking media as reliable. Consequently, mechanizing counterfeit news recognition is elementary to keep up hearty online media and informal organization. This paper proposes a model for recognizing forged news messages from twitter posts, by figuring out how to anticipate precision appraisals, in view of computerizing forged news identification in Twitter datasets. Afterwards, we performed a comparison between five well-known Machine Learning algorithms, like Support Vector Machine, Naïve Bayes Method, Logistic Regression and Recurrent Neural Network models, separately to demonstrate the efficiency of the classification performance on the dataset. Our experimental result showed that SVM and Naïve Bayes classifier outperforms the other algorithms.

Comparison of Various Machine Learning Modelsfor Accurate Detection of Fake News, Karishnu Poddar, Geraldine Bessie Amali D, Umadevi K S[6]. Fake news consists of news that is not well re-searched or deliberate steps have been taken to spread mis-information or hoaxes via different forms of news distribution networks. This paper aims to tackle this issue using a computational model of probabilistic and geometric machine learning models. Moreover, the scores of two different vectorizers namely count and Term Frequency Inverse Document Format (TF-IDF) will be compared to find the appropriate vectorizer for fake news detection. English stop words have been used to improve the scores. Various classifiers like Naive Bayes, Support Vector Machine (SVM), Logistic regression and decision tree classifier were used to predict the fake news. Simulation results indicate Support Vector Machine (SVM) with the TF-IDF gave the most accurate prediction.

3. HARDWARE AND SOFTWARE

3.1 HARDWARE

- Storage 5GB
- CPU: Intel Core i5-8400

- Memory: 8GB

3.2 SOFTWARE

- OS Version: Windows 10 64-bit
- IDE Used: Visual Studio Code
- Language: Python

4. METHODOLOGY

4.1 Data Preprocessing

There are some exploratory data analyses is performed on training data to prepare the data for modelling of system like null or missing values, removing social media slangs, removing stop-words, correcting contraction. Also, Part of Speech (PoS) Tagging has been performed in the data to meet the accuracy of prediction model. Data has been also lemmatized to get root form of the words so that prediction algorithm gets trained on the quality data. Before model training the data was tokenize so that each word in the sentence can treated as element for model training.

Lemmatization:

Lemmatization is one of the most common text pre-processing techniques used in Natural Language Processing (NLP) and machine learning in general. lemmatization involves deriving the meaning of a word from something like a dictionary. Lemmatization gives the root form of the word i.e., studying is studies. This makes sure that the root word is not just achieved by removing the suffix from a given word that is done in stemming. This makes the lemmatization algorithm slow but for the NLP technique where meaning each word is equally important by its root word. Thus, we have used lemmatization to get the root word.

Tokenization

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. Here, tokens can be either words, characters, or sub-words as tokens are the building blocks of Natural Language, the most common way of processing the raw text happens at the token level. Hence, Tokenization is the foremost step while modelling text data. Tokenization is performed on the corpus to obtain tokens. The following tokens are then used to prepare a vocabulary. Vocabulary refers to the set of unique tokens in the corpus. Remember that vocabulary can be constructed by considering each unique token in the corpus or by considering the top K Frequently Occurring Words.

4.2 Feature Selection

In this module we have performed feature selection methods from sci-kit learn python libraries. For feature selection, we have used methods like count Vectorization term frequency like tf-idf weighting.

Count Vectorization:

Vectorization is a process of converting the text data into a machine-readable form. The words are represented as vectors. we cannot pass text directly to train our models in Natural Language Processing, thus we need to convert it into numbers, which machine can understand and can perform the required modelling on it. Count Vectorizer tokenizes (tokenization means breaking down a sentence or paragraph or any text into words) the text along with performing very basic pre-processing like removing the punctuation marks, converting all the words to lowercase, etc. The vocabulary of known words is formed which is also used for encoding unseen text later. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document.

TF-IDF

TF-IDF stands for "Term Frequency — Inverse Document Frequency". This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining. TF is individual to each document and word, hence we can formulate TF as follows.

$$\text{Tf-idf} = \frac{count\ of\ t\ in\ d}{number\ of\ words\ in\ d}$$

This measures the importance of document in whole set of corpus, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N.

df(t) = occurrence of t in documents

IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

Tf-idf(t) = N/df

4.3 MODEL TRANING

In this module the extracted features are fed into different classifiers. We have used Logistic Regression, Multinomial Naive Bayes from sci-kit learn. This computational tool uses two different classifiers so that user could get more accurate results the prediction of both the classifier are shown the output page.

Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X.

$$LR(z) = \frac{1}{1+e^{z}}$$

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. As per feature selection used for the data set here the best threshold value for logistic Regression is 0.6.

Here the 80% data was used for training and 20% data was used for testing on the logistic regression classifier it gives us mean score of 0.93 and best score of 0.94.

Multinomial Naive Bayes

Multinomial Naïve Bayes uses the bag of words approach, where the individual words in the document constitute its features, and the order of the words is ignored. This technique is different from the way we communicate with each other. It treats the language like it's just a bag full of words and each message is a random handful of them.

$$P\left(\frac{B}{A}\right) = \frac{P(A\cap B)}{P(A)}$$

Naive Bayes is based on Bayes' theorem, where the adjective Naïve says that features in the dataset are mutually independent. Occurrence of one feature does not affect the probability of occurrence of the other feature. For small sample sizes, Naïve Bayes can outperform the most powerful alternatives. Being relatively robust, easy to implement, fast, and accurate, it is used in many different fields.

Multinomial NB classifier uses the 80% of data for training and 20% of data for testing with best alpha as 0.01 and gives us the mean score of 0.85 and best score of 0.93.

Sentiments Analysis

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker

VADER (Valence Aware Dictionary and sentiments Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative. VADER not only talks about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

Sentiment analysis also is used to monitor and analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

The authors can acknowledge any person/authorities in this section. This is not mandatory.

4.4 FLOWCHART

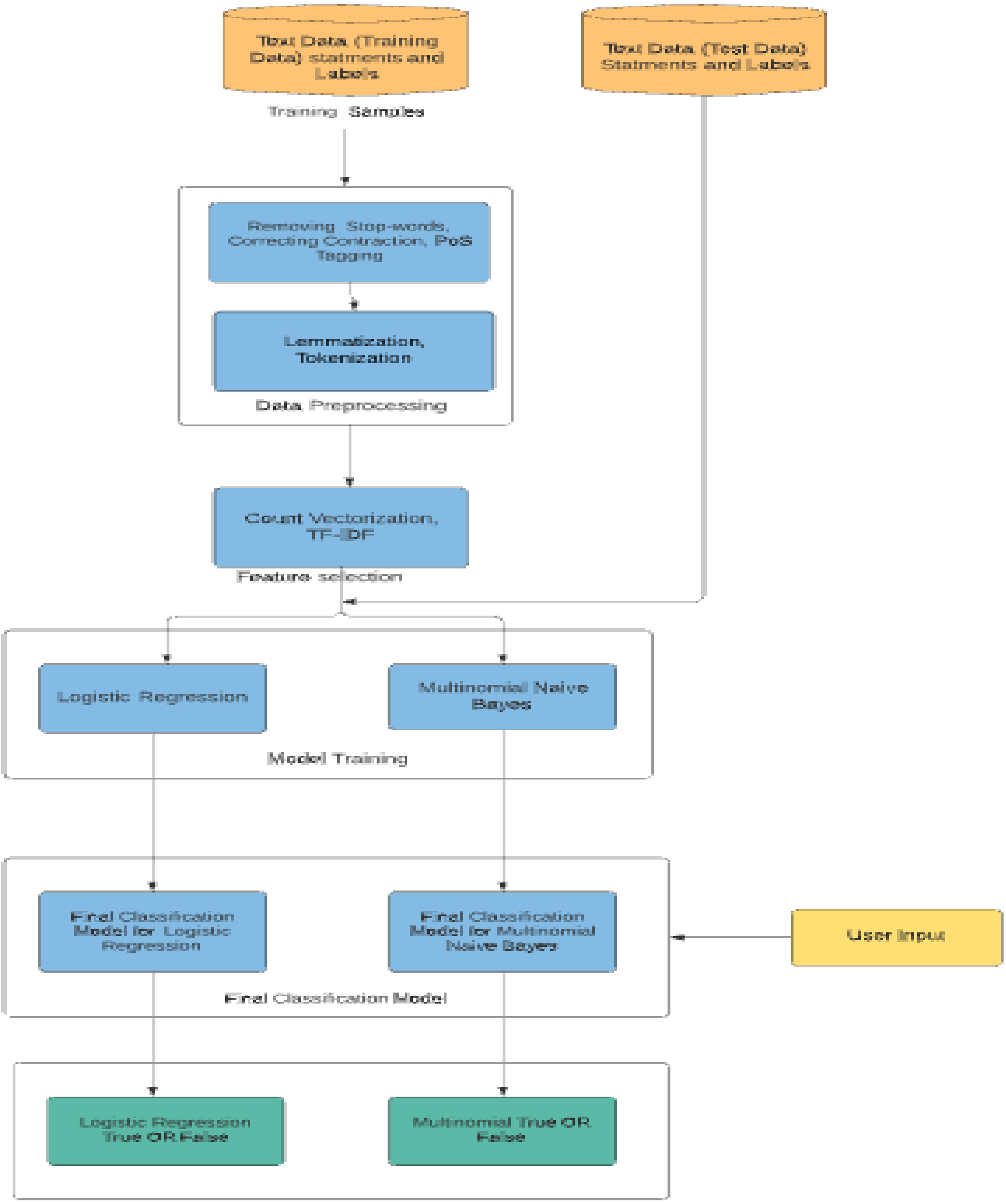

Text Data (Training Data) statments and Labels
Training Samples
Text Data (Test Data) Statments and Labels
Removing Stop-words, Correcting Contraction, PoS Tagging
Lemmatization, Tokenization
Data Preprocessing
Count Vectorization, TF-IDF
Feature selection
Logistic Regression
Multinomial Naive Bayes
Model Training
Final Classification Model for Logistic Regression
Final Classification Model for Multinomial Naive Bayes
Final Classification Model
User Input
Logistic Regression True OR False
Multinomial True OR False

Fig 4.1 Flow Chart of proposed Model

4.4.1. Data Pre-processing

First step is to perform data pre-processing on training data to prepare the data for modelling of system. Pre-processing involves steps like removing null or missing values from data set, removing social media slangs, removing stop-words, correcting contraction.

Now we will perform

1. Lemmatization on data set (To get the root form the word)
2. Tokenization.

4.4.2. Feature Selection

In this module we have performed feature selection methods from sci-kit learn python libraries. For feature selection, we have used methods like count Vectorization term frequency like TF-IDF weighting.

1. TF-IDF (compute a weight to each word which signifies the importance of the word in the document and corpus)
2. Count Vectorization (Vectorization is a process of converting the text data into a machine-readable form).

4.4.3 MODEL TRAINING

We have used Logistic Regression, Multinomial Naive Bayes from sci-kit learn.

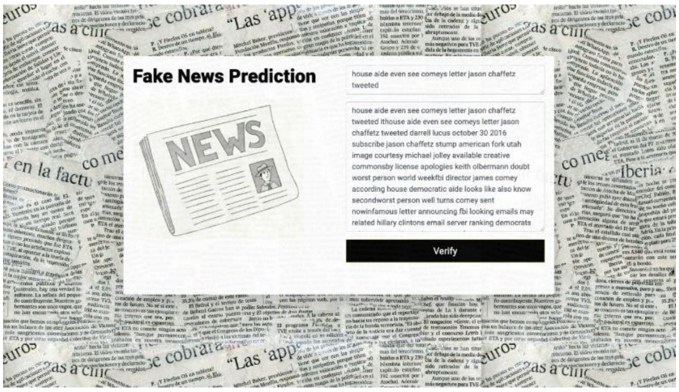
1. Training with Logistic Regression.
2. Training with Multinomial Naive Bayes.

4.4.4 TESTING

Once the model has been trained the testing needs to be done.

1. Logistic Regression uses the 80% data was used for training and 20% data was used for testing on the logistic regression classifier it gives us mean score of 0.93 and best score of 0.94.
2. Multinomial NB classifier uses the 80% of data for training and 20% of data for testing with best alpha as 0.01 and gives us the mean score of 0.85 and best score of 0.93

4.4.5 OUTPUT



house aide even see comeys letter jason chaffetz tweeted

house aide even see comeys letter jason chaffetz tweeted ithouse aide even see comeys letter jason chaffetz tweeted darrell lucus october 30 2016 subscribe jason chaffetz stump american fork utah image courtesy michael jolley available creative commonsby license apologies keith olbermann doubt worst person world weekfbi director james comey according house democratic aide looks like also know secondworst person well turns comey sent nowinfamous letter announcing fbi looking emails may related hillary clintons email server ranking democrats

Fig. 4.2 Input Screen

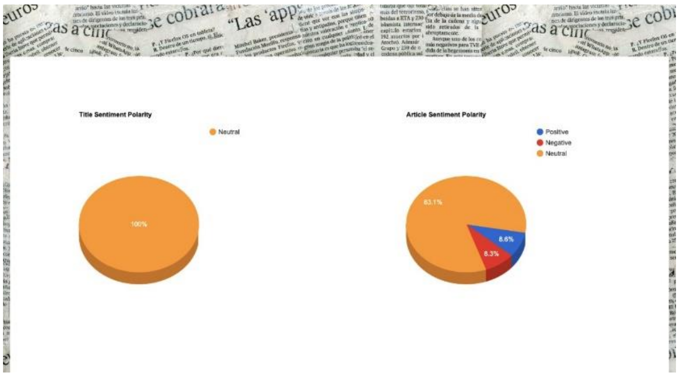


Fig. 4.3 Output Page (Sentiment Analysis)

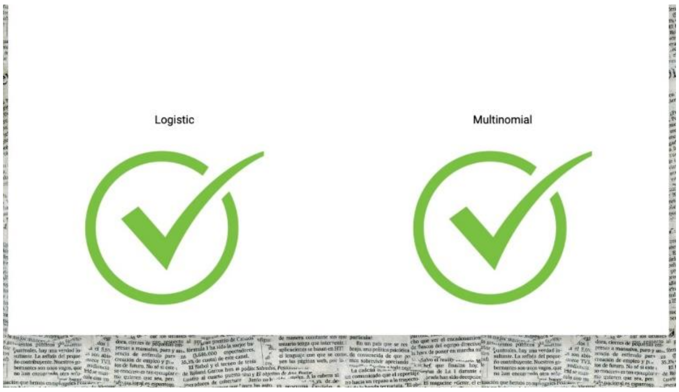


Fig 4.4 Output Page (Prediction)

5. CONCLUSION

In this paper, we've used Logistic Regression and Multinomial Naïve Bayes classifier which will predict the truthfulness of user input news, here we have presented a prediction model with feature selection used as Count Vectorization, TF-IDF which helps the model to be more accurate.

We have investigated different classifier model with feature extraction as Count Vectorization, TF-IDF. The proposed Logistic Regression model achieves the mean accuracy of 0.93 with alpha value set as 0.6.

Also, with Multinomial Naïve Bayes Model we have achieved the mean accuracy of 0.85 with alpha value 0.01.

6. FUTURE SCOPE

Future Scope of paper is vast as use of internet is increasing day by day and sharing of article, news via social media, messaging platform is increasing at vast level. To control such incidents this model could be used which shows the user about truthfulness of the message or post they intend to share to others. This will help society to avoid fake news.

7. REFERENCES

[1] Akshay Jain, Amey Kasbe[1], "Fake News Detection", The Institute of Electrical and Electronics Engineers, Published 2018

[2] Kelly Stahl[2] "Fake news detection in social media", California State University Stanislaus, 2018.

[3] Subhadra Gurav, Swati Sase, Supriya Shinde, Prachi Wabale, Sumit Hirve[3], Survey on Automated System for Fake News Detection using NLP & Machine Learning Approach, International Research Journal of Engineering and Technology (IRJET), 2019.

[4] Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita[4], Fake News Detection Using Machine Learning approaches: A systematic Review, The Institute of Electrical and Electronics Engineers, Published 2019.

[5] Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq[5], Detecting Fake News using Machine Learning and Deep Learning Algorithms, The Institute of Electrical and Electronics Engineers, Published 2019.

[6] Karishnu Poddar, Geraldine Bessie Amali D, Umadevi K S[6], Comparison of Various Machine Learning Modelsfor Accurate Detection of Fake News, The Institute of Electrical and Electronics Engineers, Published 2019

[7] Guohou Shan, James Foulds, Shimei Pan, Causal Feature Selection With Dimension Reduction For Interpretable Text Classification, University of Maryland, Baltimore County 2020