

Classifying Fake News Articles using Natural Language Processing and Supervised Learning Estimator

M Geetha Yadav¹, Rajasekhar Nennuri², N.Sairam³, Y.Shiva Teja⁴, Ganga Prasad⁵

^{1,2,3,4,5}Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad - 500043

geethayadav22@gmail.com¹, rajasekharnennuri@gmail.com², sairam.neela07@gmail.com³, shiva64462@gmail.com⁴, chilukagangaprasad@gmail.com⁵

Abstract:

In the modern days, as the internet is present everywhere, each and every one depends on variety of online sources for news. As the usage of Facebook, Twitter and many social media platforms, spreading of news is increasing rapidly among millions of people in a very short time span. Initially, the distribution of fake news spreads across the social media platforms and later finds its ways and reaches onto media such as Traditional Television and radio news. In this paper, the results of the fake news identification is studied and presented. Natural language processing, Naïve Bayes classifier or algorithm and SciPy tools are used to identify whether the news is real or fake by extracting the quotes from the news articles and estimate the likelihood percentage from the extracted quotes.

Keywords:

Fake News, NLTK, Natural language Processing(NLP), Machine learning, Naive Bayes classifier.

INTRODUCTION:

In today's era, most of their life's are spent only through online such as using the social media platforms like Facebook, Twitter and etc and very few will be their who try to avoid the social media platforms. Most of the fake news today spreads across the social media platforms itself rather than news papers and later finds its way on other different platforms. Once the news is uploaded across the social media platforms, it will then spread very fast and it is one of the difficult task to find the source of the news i.e., from where and who has started that news. Most of them think that since the news has spread all over, they consider it as a real news rather than fake news. The spreading of fake news will have a bad impact and effects every individual in the society. The standards on Social media platforms have been very less when discussing about the fake news because it is the only one platform which has the capability to spread the fake news word wide very quickly in a fraction of seconds. Social media not only benefit but also effect the people in one way or the other by spreading the fake news. Since, through news papers or through any physical documents the spread of fake news will not happen spontaneously rather than the social media. However, social media maintains spontaneity and is very fast to spread the news online. The false information or news sometimes spreads intentionally in the cases of financial purposes and very importantly in the case of politics. It has also been estimated that over millions of fake news has spread across Twitter, Facebook and other social media platforms. Many scientists believe that most of the fake news issue can be addressed using the Machine learning algorithms and techniques. With the help of the Natural language processing(NLP), it will process the text

and classifies the text based on the different Machine learning algorithms and techniques. One such algorithm we discussed here is the Naive Bayes algorithm. It is based on calculating the probability of the words such as counting the number of words, counting the number of verbs, number of unique words. Using the Naive Bayes classifier, calculate the probability score and identifying the fake and real news are presented and discussed in this paper.

BACKGROUND AND RELATED WORK:

Fake news has become one of the problem for each and everyone not only in the earlier days but also now. It has caused a great cause and effected every individual to every businesses in one way or the other. In earlier days, we used to don't have any technology to find out which news is the fake news and which news is the real. All they did is that, the people will manually try and find out about the news from many different sources whether It is a fake or real news. As we get the news, we don't even know about that news and without knowingly, we will directly consider it as a real news in most of the cases. So, as the days passed on, The Technology has come into affect. Due to this new emerging trends of technology, detecting of fake news has also become a bit easy when compared with the earlier days. Lets, now discuss about the fake news characteristics.

CHARACTERISTICS OF FAKE NEWS:

Fake news is shown to be detectable in many different ways. Checking about the fact is one of the way to identify and fake news debugging. Most of the text in the fake news will have mistakes in the grammars. The text in the fake news will be highlighted with different colors. They often try to highlight the text or news and affects the opinion of the readers. The text source also will not be true everytime. The URL will also be different in their case such as lower case, upper case or combination of both the lower and the upper case and also sometimes the URL will also contain special symbols and numerals. The attackers also try to seek attention with the help of clicking the buttons which will be highlighted and will be very much visible to the readers or viewers. All the news which shows such types of content will not always be true.

Software Requirements:

1. Programming language: Python
2. Web development : Django(HTML, CSS)
3. Software : Pandas
4. Packages used : NumPy, Scipy

Hardware Requirements:

1. Speed : 2.4GHz
2. RAM : 512 MB
3. Hard Disk : 20 GB
4. KeyBoard : Standard
5. Monitor

Proposed System:

The methodology used here after the research is using the Naive Bayes algorithm which of the one of the algorithm of supervised learning.

In this paper, the libraries of Python such as Sci-kit are used. Python is one of the open source IDE or software to download and install. Python has many and huge extensions and libraries which helps to develop with one of the emerged technology such as Machine learning. All the different types of machine learning algorithms are available in Sci-kit libraries which are available for Python. Django is one of the deployment model which can be used to develop the web pages using HTML, CSS, Javascript which will provide the client side implementation.

Process flow:

The steps involved in this procedure and methodology are:

Step 1 : Loading the Dataset.

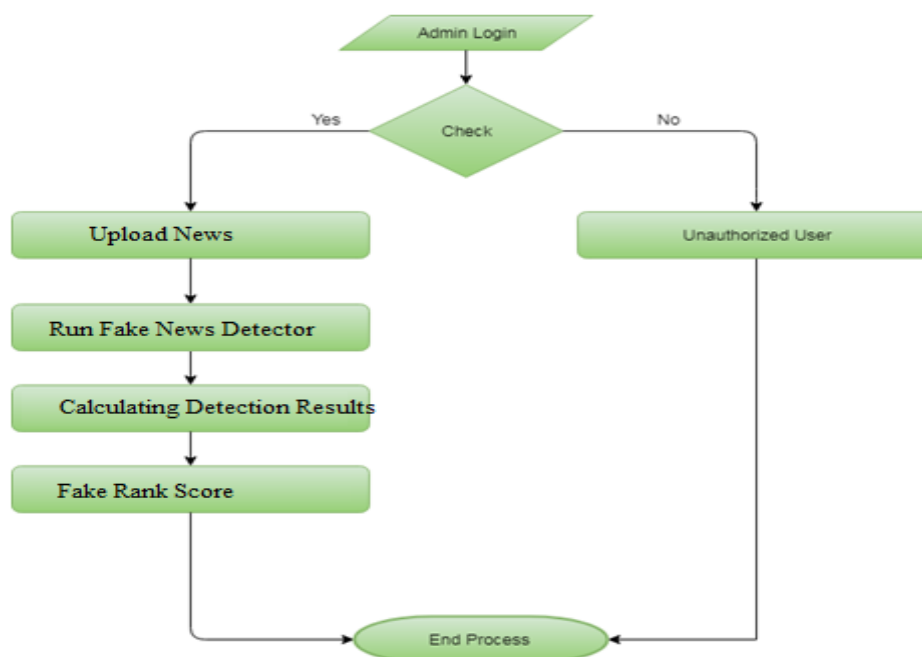
Step 2 : Pre processing the Date from the dataset such as stop words. Tokenizing and etc.

Step 3 : Feature Selection such as Count features and Tf-idf features.

Step 4 : Applying the classsication model such as Naive Bayes classifier.

Step 5 : Classifying the new data and finding the probability about the text as True or False along with the probability percentage. If the probability score is more than 0, it will be considered as fake news else real news.

Score will be calculated based on the length of the number of verbs, number of entities, number of double quotes and etc.



IMPLEMENTATION:

The Naive Bayes classifier will be used to calculate the length of the text, number of verbs, number of entities, number of quoted words such as single quoted or double quoted. The attribution classifier actually works based on 3 factors – Source, Cue and Content.

Source	The span of text that includes who put forth the quote or who the content is attributed to.
Cue	A verb or verb phrase that lexically links the source to the quote or content.
Content	The span of text that serves as the quote and is attributed.

Let C be any content span of random span length $\text{len}(C)$ for a quote which will require attribution. The span of the attribution is the absolute distance “D” in character spaces from the starting or ending of a span content marked by double quotes. So for any quote attribution, we can denote as:

$$(\text{Source}, \text{Cue}) \leq x^i + \text{len}(C) + 2D$$

(or)

$$(\text{Source}, \text{Cue}) \geq x^i$$

Cue identification is actually based on the associated verbs or information which is present within the training dataset. Most of the informative words or verbs or phrases will be added into a simple model like “bag of words”. Feature extraction from the text will be done by applying Machine learning algorithms.

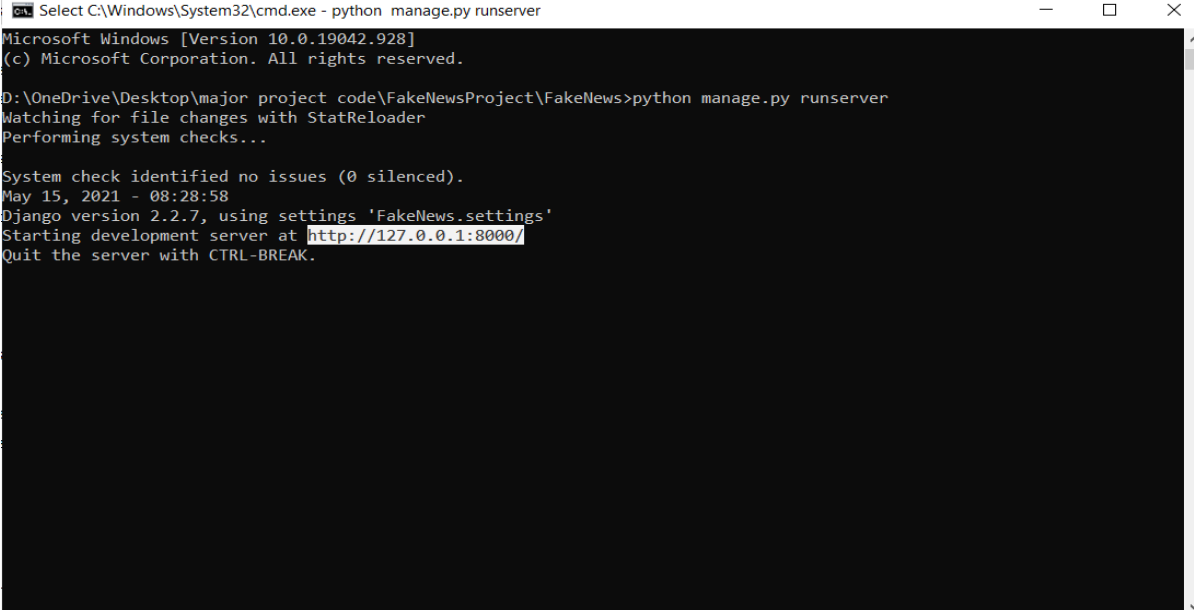
FAKE NEWS DETECTION ALGORITHM:

The tool or the classifier works based on the scoring system. The fake news detection algorithm works in this manner: The text in the document will be initially counted and tokenized. Then, the algorithm will check for the quotes such as single or double quotes in the text or the paragraph within the document. It will then remove the stop words, and the common words using the Naive Bayes algorithm. With the help of Naive Bayes classifier, it will classify the text in the document and calculate the probabilities and score. Based on the score, the text in the document will be classified as a fake or the real news. We initially use the training datasets to train the model followed by development and in the final, we test the model with the help of the testing datasets.

Algorithm :

```
1 For each doc in Documents
2 pCount  $\leftarrow \Sigma$ (number of paragraphs)
3 Tokenize doc by paragraph
4 FakeRank  $\leftarrow 0$ 
5 For each paragraph in doc
6   qScore  $\leftarrow \Sigma$  (number of quotes)
7   if qScore > 0 then
8     A-score  $\leftarrow 0$ 
9     For each quoteset
10      quote_cl classify.naivebayes(quoteset_attribution_space, d)
11      If quote_cl
12        A-score  $\leftarrow$  A-score +1
13      Else
14        A-score  $\leftarrow$  A-score -1
15      return A-score
16   FakeRank  $\leftarrow$  FakeRank + A-score
17 If FakeRank  $\geq 0$  then
18   docLabel  $\leftarrow$  real
19 If FakeRank < 0 then
20 docLabel  $\leftarrow$  fake
```

Step 1 : Run the command “python manage.py runserver”

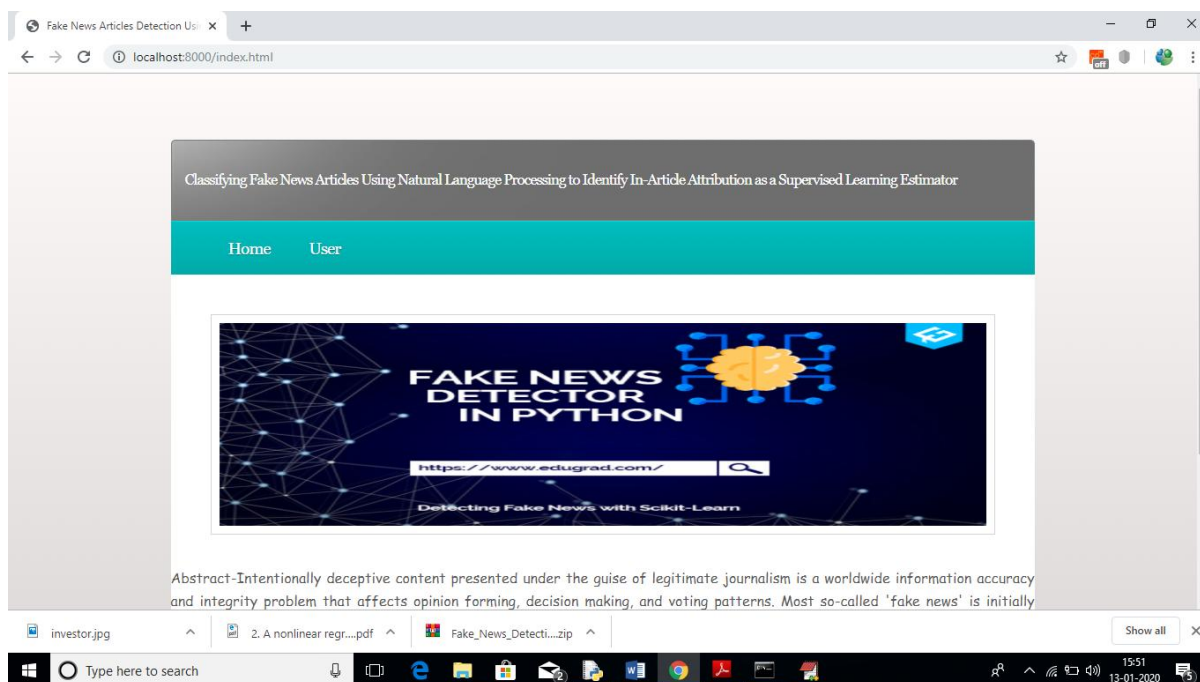


```
Select C:\Windows\System32\cmd.exe - python manage.py runserver
Microsoft Windows [Version 10.0.19042.928]
(c) Microsoft Corporation. All rights reserved.

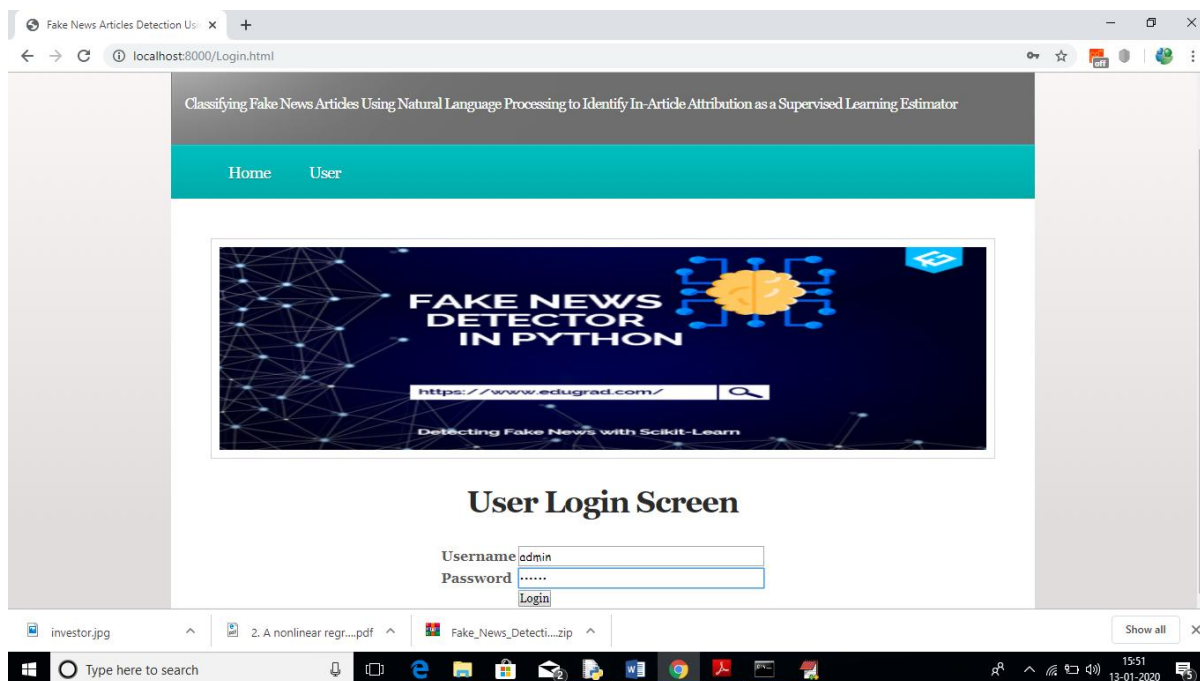
D:\OneDrive\Desktop\major project code\FakeNewsProject\FakeNews>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).
May 15, 2021 - 08:28:58
Django version 2.2.7, using settings 'FakeNews.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

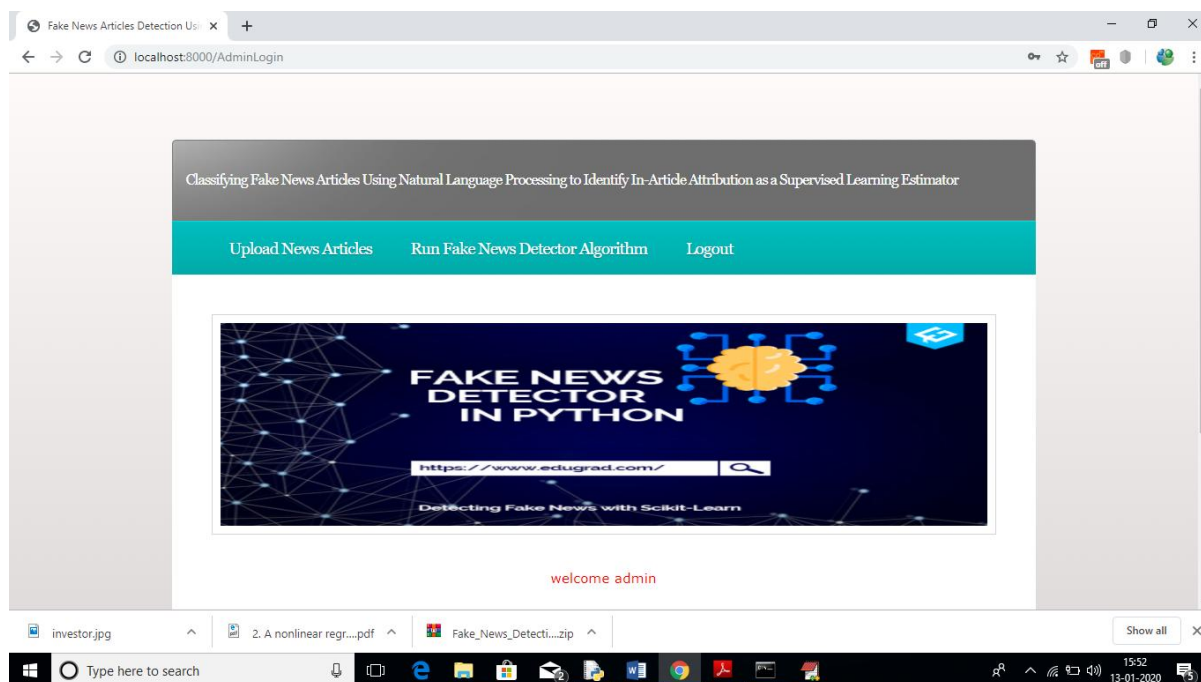
Step 2 : From the browser type <http://127.0.0.1:8000/index.html> and it will display the Home page.



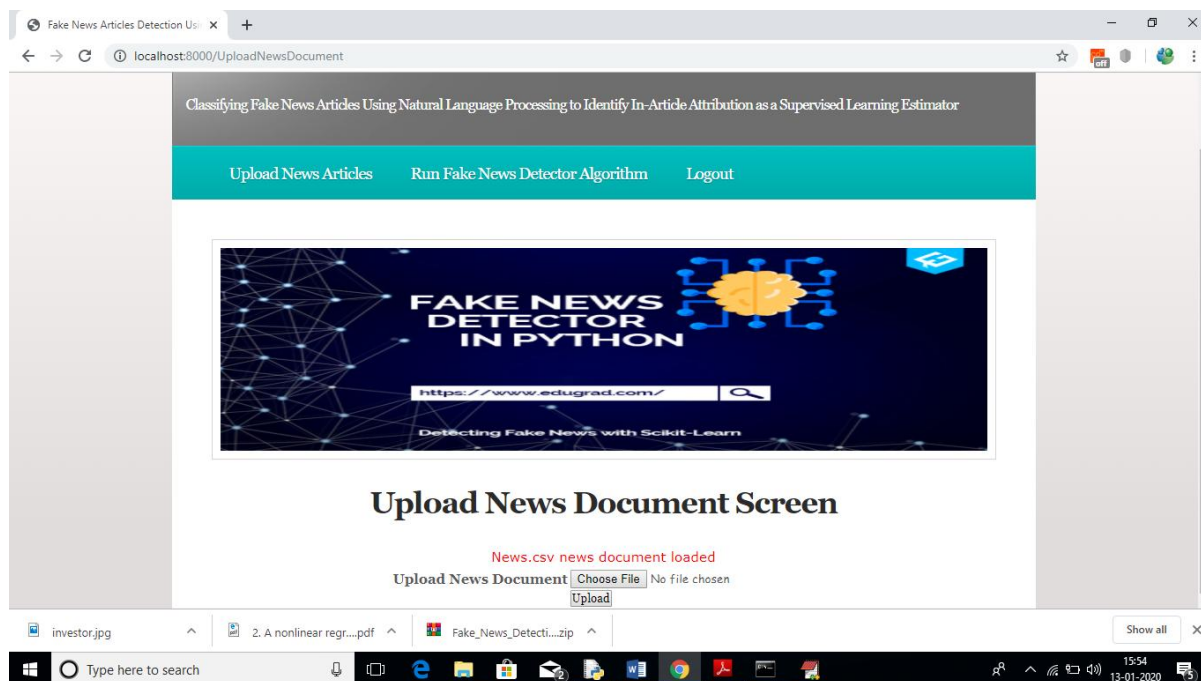
Step 3 : Click on the “User” to open the login page.



Step 4 : Give the correct credentials and then try to login.



Step 5 : Click on “Upload News Articles” and then try to upload the dataset(.csv file)



Step 6 : Then, Run Fake News Detection Algorithm and the algorithm runs and displays the results such as fake or real news and also the score.

News Text	Detection Result	Fake Rank Score
Says the Annies List political group supports third-trimester abortions on demand.	Fake News	0.8333333333333333
When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bushs administration.	Real News	2.142857142857143
"Hillary Clinton agrees with John McCain ""by voting to give George Bush the benefit of the doubt on Iran. ""	Real News	3.076923076923077
Health care reform legislation is likely to mandate free sex change surgeries.	Fake News	0.7692307692307693
The economic turnaround started at the end of my term.	Real News	0.9090909090909092
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	Real News	1.3333333333333333
Jim Dunnam has not lived in the district he represents for years now.	Real News	2.142857142857143
"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."	Real News	1.5151515151515151
"However, it took \$19.5 million in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific."	Real News	2.142857142857143
Says GOP primary opponents Glenn Grothman and Joe Leibham cast a compromise vote that cost \$788 million in higher electricity costs.	Real News	2.1739130434782608
"For the first time in history, the share of the national popular vote margin is smaller than the Latino vote margin."	Fake News	0.8
"Since 2000, nearly 12 million Americans have slipped out of the middle class and into poverty."	Real News	1.5
"When Mitt Romney was governor of Massachusetts, we didnt just slow the rate of growth of our government, we actually cut it."	Real News	2.2222222222222223
The economy bled \$24 billion due to the government shutdown.	Fake News	0.8333333333333333
Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended.	Real News	2.1052631578947367
"In this last election in November, ... 63 percent of the American people chose not to vote, ... 80 percent of young people, (and) 75 percent of low-income workers chose not to vote."	Real News	0.975609756097561

DEFINITIONS:

Pre-Processing Data

Data will be in many different formats such as structured, semi-structured and unstructured. We need to categorise the data. Then, Cleaning of the data involves various stages:

- **Remove Punctuation**

It removes the punctuation marks present in text

Eg. How do you do? → How do you do

- **Tokenization:**

Tokenizing will split the text into words.

Eg. The cat is very cute → "The", "cat", "is", "very", "cute"

- **Remove Stop words:**

Stop words can occur in any of the document containing text.

Eg. BallandBat are needed → Ball, Bat, needed

- **Stemming:**

Stemming generally removes the suffices attached to the text.

Eg. Playing, played → "ing", "ed"

Feature Generation:

Then, with the help of the text we got, we can generate few features such as count the number of words, large words frequency, and also the word frequency of the unique words.

Vectorising Data using TD-IDF:

TF denotes the “Term Frequency” and it finds us the about the frequency of a term which has appeared in the document.

$$TF(t, d) = \frac{\text{Number of times } t \text{ occurs in document 'd'}}{\text{Total word count of document 'd'}}$$

IDF stands for ‘Inverse Document Frequency’. IDF removes most of the common words which appear most of the times in a dataset like “a”, “an”, “the”, “on”, “of” etc. IDF is actually used to bring down the term importance of the common words and the rare terms will be given more importance.

$$IDF(t, d) = \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

Classification using Naive Bayes classifier:

Naive Bayes is a supervised learning algorithm which is based on Bayes theorem and used in solving classification problems. It is a probabilistic classifier, which means it predicts on the basis of probability by calculating the number of words, verb count and etc.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Converting the dataset into frequency tables. Then, generating the likelihood table by finding the probabilities of the features given. Finally, calculate the probability.

CONCLUSION:

This paper presents the results of the detection of fake news using this algorithm in an efficient way. Now a days, spreading of fake news has become one of the biggest issue which we are facing. We are so confused about any news we get and also becoming a bit tough to identify whether the news we get is the fake or the real news. So, this fake news Detection system has been developed to eradicate the issues faced by the fake news. This system will take the input from the user and classifies it as a real or fake news and also gives the percentage of the detected news. To implement this system, we have used Natural language

Processing(NLP) and Machine learning algorithms such as Naive Baye's classifier. The model will be trained initially by taking the use of the appropriate training datasets and therefore the model is then tested using the appropriate testing datasets. Based on the probability and the score calculated from the algorithm, the news will be considered as a fake or the real news.

REFERENCES:

- [1] M. Balmas, "When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism," *Communic. Res.*, vol. 41, no. 3, pp. 430–454, 2014.
- [2] C. Silverman and J. Singer-Vine, "Most Americans Who See Fake News Believe It, New Survey Says," *BuzzFeed News*, 06-Dec-2016.
- [3] P. R. Brewer, D. G. Young, and M. Morreale, "The Impact of Real News about "Fake News": Intertextual Processes and Political Satire," *Int. J. Public Opin. Res.*, vol. 25, no. 3, 2013.
- [4] D. Berkowitz and D. A. Schwartz, "Miley, CNN and The Onion," *Journal. Pract.*, vol. 10, no. 1, pp. 1–17, Jan. 2016.
- [5] C. Kang, "Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking," *New York Times*, 21-Nov-2016.
- [6] C. Kang and A. Goldman, "In Washington Pizzeria Attack, Fake News Brought Real Guns," *New York Times*, 05-Dec-2016.
- [7] MG Yadav, R Nennuri, "Data Mining based Modern and Advanced Design and Development Applications", *International Journal of Pure and Applied Mathematics* 119 (16), 4651-4658
- [8] R Nennuri, AK Chaitanya, LP Malyala, "Implementation of data frame work system based on model driven architecture for MAS and Web based applications", *International Journal of Engineering & Technology* 7 (2.20), 1-4
- [9] Rajasekhar Nennuri , Prathyusha Malyala , SaiTejaswi Thotakura, "Crime Prediction and Analysis Using Machine Learning", *International Journal of Advanced Science and Technology* 29 (special issue), 9
- [10] YB M Geetha yadav Golla Swetha , Vasista Kumar, "Intrusion Detection Scheme Using Machine Learning", *icrcsit-20*
- [11] KS M Geetha Yadav , S Srihitha , C Tejaswi, "Clustering of Bigdata in Application Review Analysis", *International Journal of Advanced Science and Technology* 29 (special issue)

Reproduced with permission of copyright owner. Further reproduction
prohibited without permission.