Fake News Detection on Social Media

Abstract

Social media becomes very dangerous when it comes to news consumption. If we see the both sides, we can see that it has both negative and positive aspects. The positive aspects are its low cost, easy to access and the increase in the gain of information enables the consumer to share the information with others. On the other hand, its negative aspect is that it can spread the fake news which is low quality news which contain fake information. Which could do a lot of harm to an individual or the whole society. For Example, in the elections of 2016 in United States the fake news spread faster than the realistic and authentic news which increases the need and attention towards detection the false news.

The goals of this application are to assist people and save them time when determining if a news article is fake or not. They don't have to go through the process of manually examining it. Fake news and rumors are spreading at an increasing rate, reaching an increasing number of individuals and entering deeper into social media. Our goal is to create a trustworthy model that can determine whether a news article is fake or not.

Chapter 1 Introduction

1.1. Introduction

With the blast of the internet with its world and the fastest increase in its popularity and usage, our world sees a boom of knowledge in mankind's history. With its popularity, every organization has taken a lot of advantages like making their employees informed from every possible news and knowledge. The past methods of delivering news took a large hit badly and got in the back because of these social media platforms and other modern news platforms. In their current situation, these social networking sites are extremely powerful and useful because they help individuals to debate and trade thoughts, as well as discuss topics. Certain entities, on the other hand, use such channels for harmful purposes such as creating prejudiced ideas, manipulating mindsets, and disseminating satire or ridiculousness, among other things. This phenomenon is known as fake news.

In the past few years, research shows a tremendous amount of surge in the propagation of false information/news, for example, in the election conducted by the US in 2016 [1]. This type of growth in publishing misleading articles which didn't conform to the reality of truth causes more complex problems in every field of life. The human brain works on what information it sees and according to that information it performs actions and these new era information platforms are the modern way of getting the information. There is a flaw in these platforms is that not have verified information. The nature of a human is usually is that what his eyes see on these platforms, he directly accepted and adopted that information, and if the information is proved wrong later on then the mind feels shocked and behaves in a rather bad and very crazy manner.

Researchers maintain a variety of repositories that have/includes the records of websites that are declared questionable or false. But these services require human knowledge to distinguish between bogus or fake websites or articles. Furthermore, truth-checker online sites only distinguish between specific areas, such as diplomacy, but they are not able to recognize rumors or false information from a wide range of areas like entertainment, sporting events or innovations, etc. There are a variety of formats the data is uploaded on the online world, for example, in the form of a text file, a video clip file, or an audio file, it is difficult to recognize and classify content posted publicly in an informal manner like news articles, audio files, or video clips Because it requires intelligence like humans to identify it [2]. Many techniques like Natural language processing (NLP) can detect irregularities that distinguish deceptive or misleading information or news from the information based on facts.

These kinds of misleading information are divided into numerous forms on a conceptual level. This factual knowledge is used to produce different ML models throughout various areas, such as retrieving different characteristics like training of various Artificial Intelligence-based Models. The research shows that the validity of an article will drop if the n-grams grow. This case is used for the classification of the ML models. We can enhance the accuracies by merging textual characterization with additional information using different ML models.

1.2. Overall Description

The overall description of the project is as follows: -

1.2.1. Objectives

The goals of this application are to assist people and save them time when determining if a news article is fake or not. They don't have to go through the process of manually examining it. Fake news and rumors are spreading at an increasing rate, reaching an increasing number of individuals and entering deeper into social media. Our goal is to create a trustworthy model that can accurately check that whether a given news article is true or not.

1.2.2. Problem Description

In the recent decade, with the blast of internet with its world the false informations numbers are increasing rapidly. Which is the main cause of misleading and performing of bad actions by the people who are infacted by this fake information. So, it's the need of a day to develop an automated system that can distinguish between "fake" and "real" news articles after being trained using a certain dataset.

1.2.3. Methodology

A lot of research has been done on identifying and categorizing misleading information (fake news) on online sites/platforms. These kinds of misleading information are divided into numerous forms on a conceptual level. This factual knowledge is used to produce different ML models throughout various areas, such as retrieving different characteristics like training of various Machine learning Models such as Support Vector Machine (SVM), Logistic Regression (LR), etc. The research shows that the validity of an article will drop if the n-grams grow. This case is used for the classification of the ML models. We can enhance the accuracies by merging textual characterization with additional information using different ML models. Recurrent Neural networks (RNN) and long-term short Memory (LSTM) of deep learning are also used to enhance accuracy. We'll code with Anaconda Software, and we'll utilize an English fake news articles dataset to train and test the machine.

1.2.4. Product Scope

We'll create an algorithm to aid social media platforms in detecting false news articles written in English. This will save you time while also providing accurate findings. We no longer need to manually verify the news article, wasting your time; instead, our system will do it for you. It will verify the authenticity of any news article and determine if it is false or genuine.

1.2.5. Operating Environment

The system will operate on the following specification

- **Minimum Requirements**

OS: Windows 7 or above

Processor: Intel Core i5 3rd generation processor

Memory: 4 GB or above

1.3. Assumptions and Dependencies

A dataset for good recognition of system to its environment. Text Articles used, after which an automated computational technique can be employed to complete the necessary analysis and interpretation.

1.4. External Interface Requirements

The external interface requirements are as follow:

1.4.1. Hardware

- PC computers, Laptop, Windows Machine

1.4.2. Software

Fake news detection is a machine learning model. The following tool will be used in the development of the algorithm:

- Anaconda 3 2020.11(64 bit) or latter
- Jupiter notebook
- PyCharm

1.5. System Features

We proposed a Machine learning-based method for this project. The fact that this algorithm can learn the features in a hierarchical manner means that the learned features are more discriminative and concise than the features that were manually created.

1.5.1. Description and Priority

It will choose and load text articles from the source. The feature will be extracted, and the news article will be detected as fake or real.

1.5.2. Response Sequences

When the text article load, the system will indicate whether the article is fake or not.

Chapter 2 Literature Review

2.1 Overview

Internet and its informations plateforms becomes very dangerous when it comes to news consumption. If we see both sides, we can see that it has both negative and positive aspects. The positive aspects are its low cost, ease to access and the increase in the gain of information enables the consumer to share the information with others. On the other hand, its negative aspect is that it can spread fake news which is low-quality news that contains fake information. Which could do a lot of harm to an individual or the whole society. For Example, in the elections of 2016 in the United States, fake news spread faster than realistic and authentic news which increases the need and attention for detection the false news [3].

The fake news detection opens up difficult and new challenges. For example, it is very difficult to detect the false information written to misguide readers for making them believing in the fake information. So, we must add some additional information to help the consumer to distinguish between fake news and real news.

There are many other methods which are used to detect fake news, articles based upon their content and these methods are based on truth-checker websites. There are variety of records maintain which contain the information about the websites which are questionable or false. However, to maintain this data requires a lot of human effort and knowledge to differentiate between fake or real sites. But these sites only differentiate in their specific domain like if the sites are distinguishing the news articles about educational system, then it will only distinguish the new article on education not on other domains.

The False news has its own unique qualities. The Malicious accounts can easily and quickly create to spread fake news [3]. The social media consumers are exposed to many varieties of news on the media platforms which results in many possibilities like the same type of people can make a group the media platform which create the environment where the consumer only encounters information or opinions that reflect and reinforce their own.

2.2 Related Works

There are many methods tried by many researchers to solve the problem of fake news to check which produces the best fit outcome and results. In the few research publications feature extraction and model creation techniques using data mining processes are considered for the detection of fake news. The combination of feature extraction (both news content and social context features) and metric evaluation (precision, recall, and f1 scores) has yielded educated findings, but the challenge is not that straightforward.

Other factors influencing the forecasts include bot spamming, click bait, and news source. These were some data mining and natural language processing-focused approaches, but as AI research and development progressed, researchers became more interested in heavier neural network-based approaches [4]. An article described a method for detecting fake news by using recurrent neural networks to "collect," "score," "integrate," and "build" a model. They employed a recurrent neural network to capture the temporal pattern of user activity around a certain article/text, and then extracted source attributes based on the user behavior. All of this information is combined to create a model for identifying fake news.

The research shows that even a simple network model can outperform more complicated models. Clearly, model complexity is not the best approach here; rather, the proper selection of parameters and data is critical. Some have tried linguistically infused neural networks, while others have employed Convolutional neural networks. In radionics, the term "DL" is used. A CNN model was applied for histological categorization of head and neck malignancies by Ye J. et al, and the accuracy, sensitivity, and specificity were all 0.79, 0.71, and 0.85, respectively [5].

Another study uses a linguistically infused neural network model using Long-Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) to classify tweets. So, while numerous attempts have been made, everything is a little jumbled and dispersed. There is a lot of room for research and growth in this field, especially because news statements have so many variables: sarcasm, abbreviation, metaphors, and so on. However, attempts have been undertaken to compile a quality dataset from dependable and large data. This project made use of one such benchmark dataset.

The most difficult task of all is how to classifiy the false news or information from the real one. But the researches have shoen that AI-based models are effecitive and the solution to this problem like model of ML, DL etc. These AI-based models are very effective in classifying the false info from the real one. In todays world there are thousands of dataset available online which can used for testing and training purposes. Many researches have shown that the AI-based ML models more effective then others and gives more accurated result

2.2.1 Terminology

Following are the terminologies used in this project:

- Natural Language Processing (NLP)
- **Machine Learning Models**
- ➢ Logistic Regression (LR)
- ➢ Support Vector Machine (SVM)
- ➢ Naive Bayes
- **Bert**
- **Deep Learning Models**
- ➢ Recurrent Neural Network (RNN) and Long-Short-Term Memory (LSTM)
- **Domain:** Machine Learning & NLP
- **Dataset**: Fake New Articles Datasets
- Text Articles

2.2.2 Categorization of Existing Techniques/Works/Research

Existing strategies, which I discovered have mostly focused on the past classification-based machine learning methods to enhance the accuracy, were applied to certain text or articles or images in an effort to address the detection of fake news. However, they did not use the technique that distinguishes our work from the other existing works.

2.2.3 Proposed Improvements in Existing Works

This sort of research has been done before, and it provides a revolutionary machine learning approach for the diagnosis of false news detection. Our strategy is primarily designed to make the life of a common person easier. When it comes to detecting the fake news, our technologies have significantly reduced the time-consummation [2]. Additionally, our technology has the potential to reduce inter-observer variability when it comes to obtaining detecting the false news articles.

2.3 Summary

According to the findings of this study NLP, Logistic Regression (LR), SVM models are used to design a method for the automated classification of true vs. false news articles. These Machine Learning models are used to develop the proposed detection model we will use and Supervised Machine Learning model. For the training of the model, the Fake new articles dataset will be used. This study demonstrates that Machine learning has the potential to be used to create a detecting classifier that may give near real-time detection of labeling for inappropriate articles detection by being totally trainable on a library of news articles taken from online sites.

Chapter 3 System Design

3.1 Introduction

This project is a research-based project which aims to build an AI-based terminology that automated the detection of fake news articles mentioned in the above chapters, this chapter is based upon the user interface and design of the applications so not much of the discussion will be involved in this chapter. We only discuss the workflow and activity diagram of the project.

3.1.1 Purpose

This chapter is written because it will help in the implementation of the project, Fake new articles detection using articles data. On the other hand, it will make a clear picture of what will be the outcome of the project.

3.1.2 System Overview

This project aims to build a model based on Machine learning techniques to detect fake news articles without human effort or without needing a human to find out whether a news article is fake or real. The project is mainly to build and train a model which can recognize a pattern of a news article for this purpose, a dataset is available that contains text articles of fake or real data. The model will take these articles as input and decide whether the article is fake or real.

3.1.3 Design Map

To build the model, real-time articles will be given to the model then the text articles will be preprocessed, and then Classification will be performed on it. After successful classification feature extraction will be performed using NLP to detect fake news. In case of fake news, the system will mark it as fake news.

3.2 Design Considerations

There is not much consideration for the design as it is a research-based project.

3.2.1 Assumptions

The project assumes that the environment is a fake news article and which are to discriminates them as fake or real.

3.2.2 Constraints

Data should be in the form of text, other forms of data would not be process-able for the model and might not be able to perform any useful task.

3.2.3 Risks and Volatile Areas

The project does not have any risks and volatile areas infect the project aims public safety.

3.2.4 Risk Mitigation

There is no risk mitigation in the project.

3.3 Activity Diagram

The below activity diagram shows the implementation process of this project and the steps of how it will work and detect the fake news and real news.

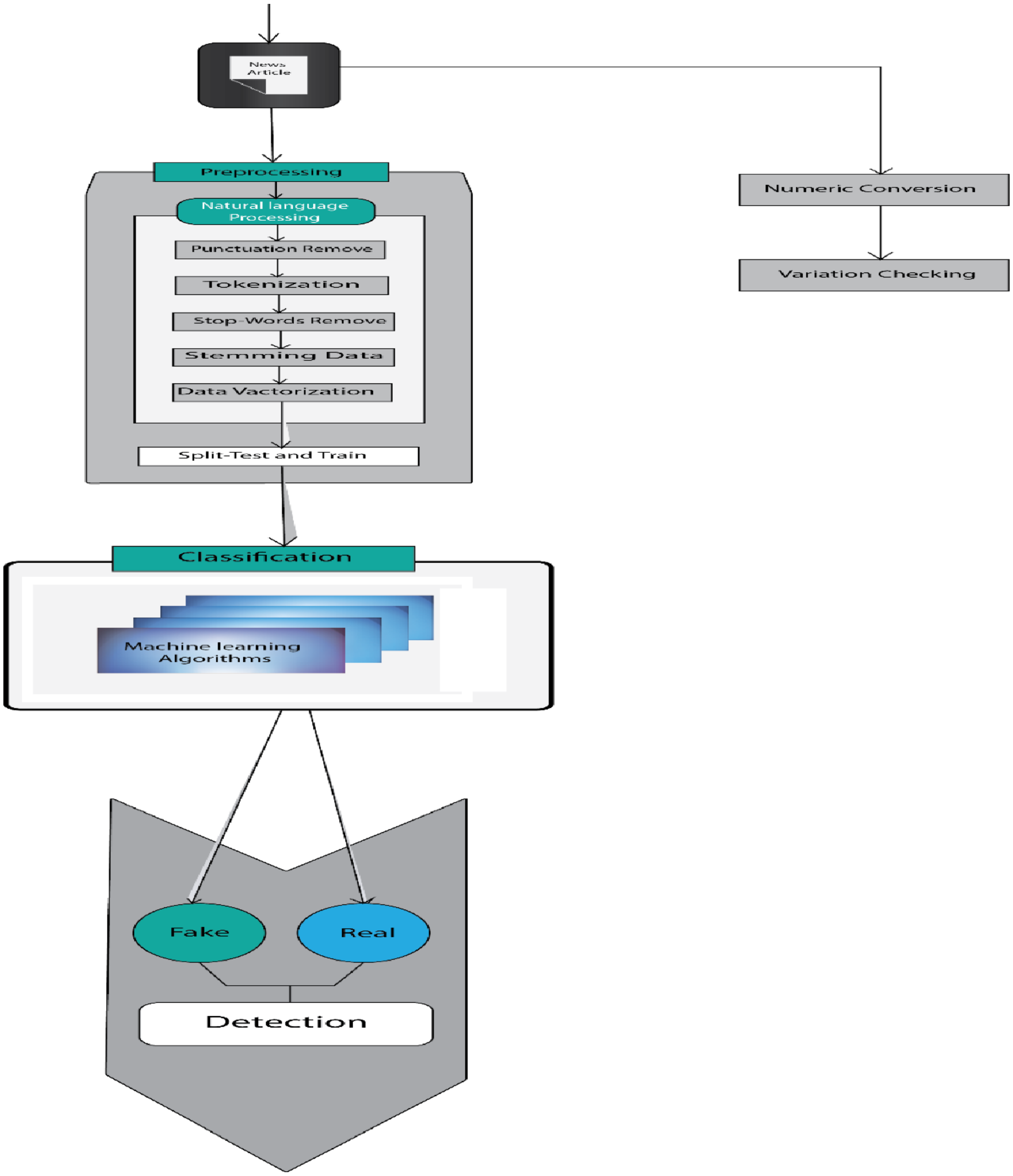


Figure 3.1 Activity Diagram

Chapter 4 Implementation

4.1. Discussion

In this project we use text dataset which has the features of the “Title of the article”, “Author of the article” and the “article text” and our dataset has 20800 training examples and 5 feature. We use NLP for the pre-processing task and then we will be applying Logistic Regression, SVM, Naive Bayes Machine Learning Classifiers for the detection of fake or real article and then we will apply LSTM and RNN to enhance the accuracy. On the first phase of accuracy of the Machine Learning models are 99%, 96% and 76%, and we are satisfied with our model and its predictions.

4.2. Development Methodologies

This is a system for detecting fake news article and it is carried out in two phases. In the first phase we are using NLP for the pre-processing of our dataset and then we use supervised Machine learning model Logistic Regression and SVM, Naive Bayes and then we apply to enhance the accuracy deep learning models LSTM and RNN, for the purposes of the classification and detection of our article is fake or real.

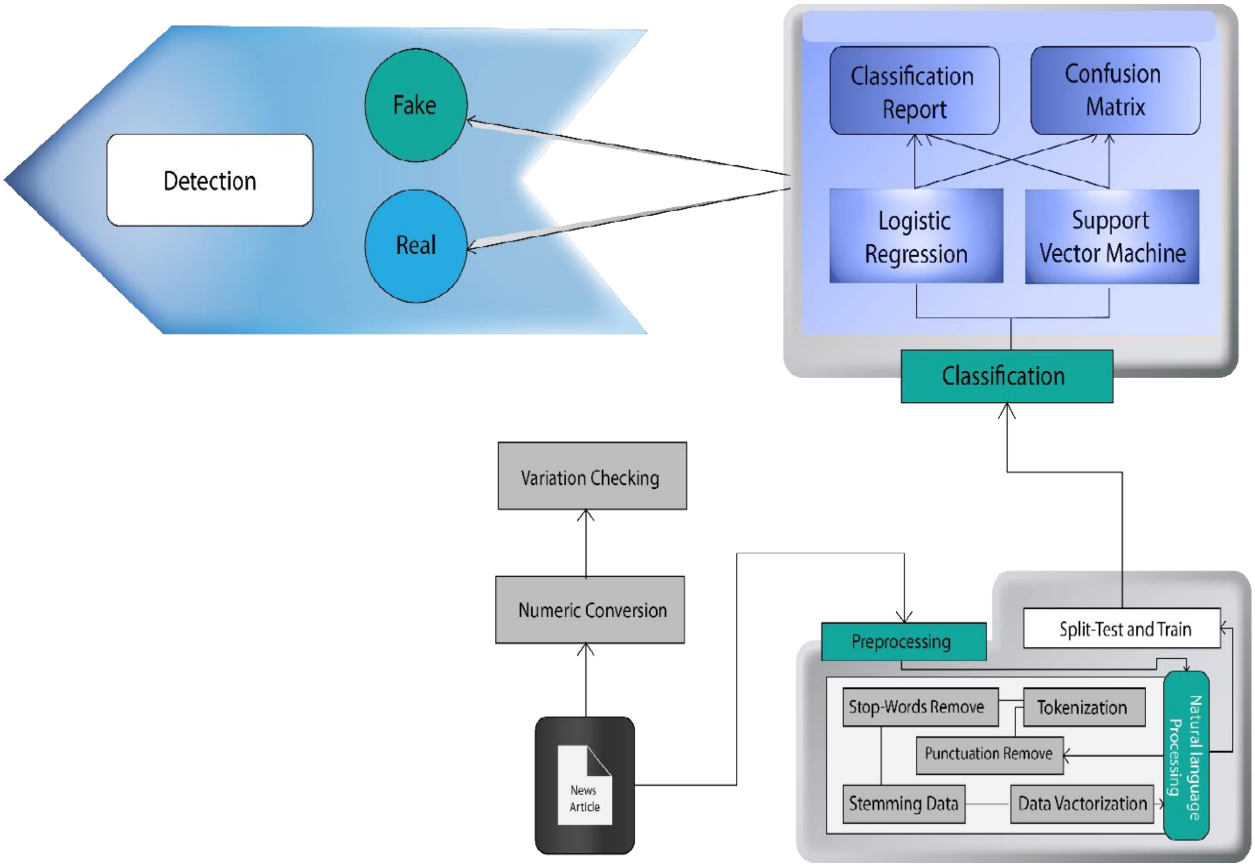


Figure 4. 1 Development Methodology

4.3. Implementation Tools and Technologies

We use python for our development programming language and a tool, we also use for development purposes. At the start of this project, we have the knowledge needed for the implementation of the project. We have expertise in python and have done multiple projects in the domain of AI. We have studied in some relevant courses related to AI, Machine Learning, and Deep Learning.

4.3.1. Technologies

Following are the technologies used in this project:

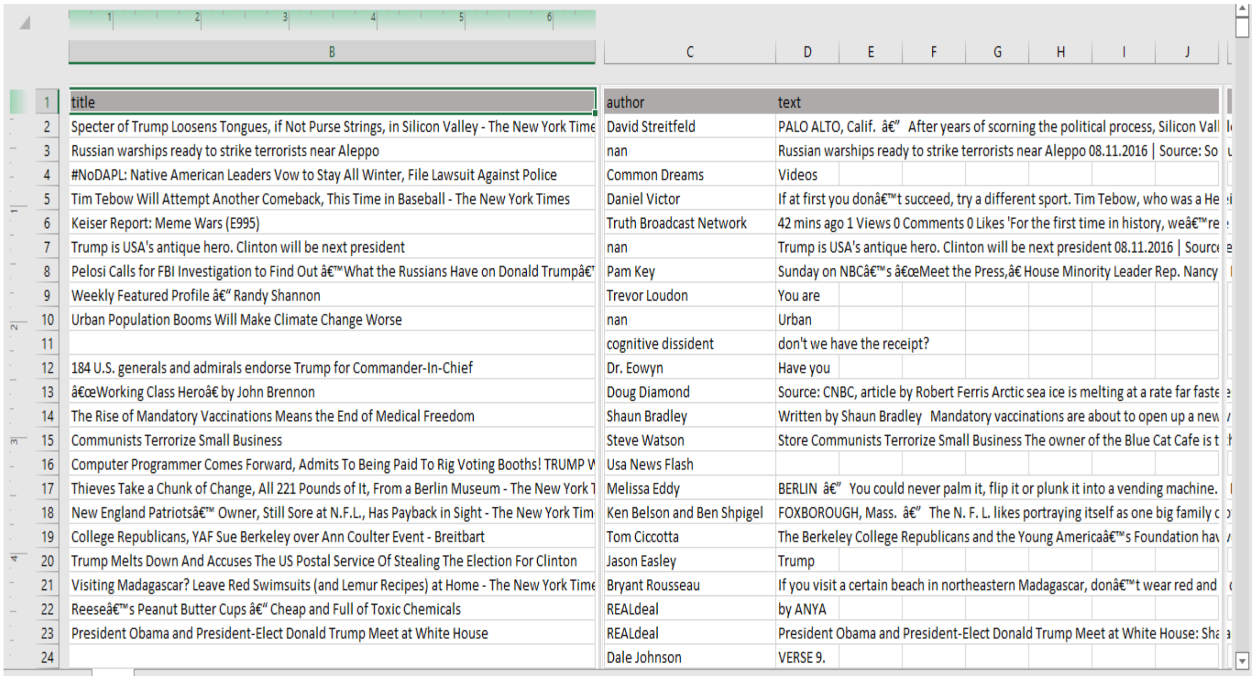
- **Hardware**
- ➢ Personal Laptops
- **Algorithms**
- ➢ Natural Language Processing (NLP)
- ➢ Logistic Regression
- ➢ Support Vector Machine (SVM)
- ➢ Naive Bayes
- ➢ Long-Short Term Memory (LSTM)
- ➢ Recurrent Neural Network (RNN)
- ➢ Bert
- **Software**
- ➢ Python
- ➢ Anaconda as IDE.

4.4. Implementation

The implementation details of this project are discussed below:

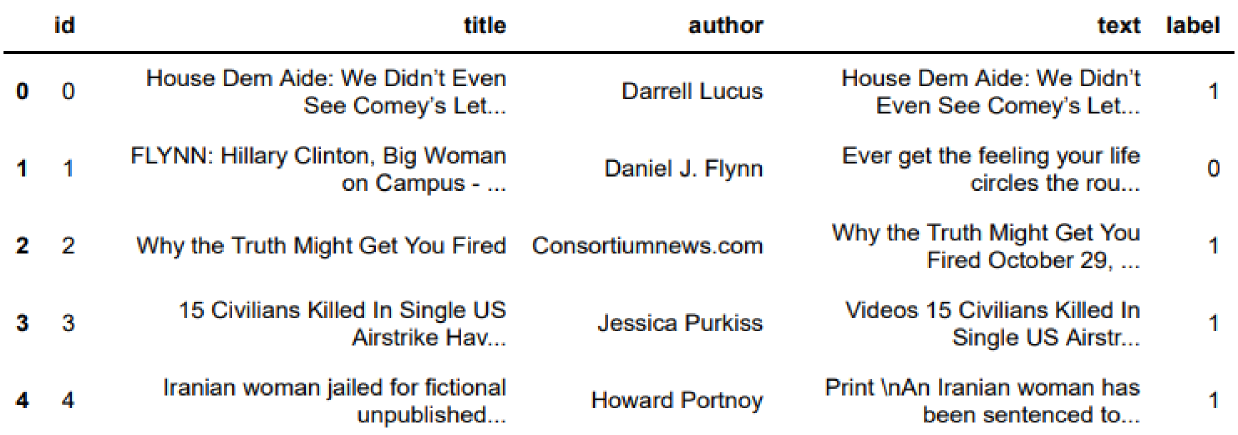
4.5. Dataset

In this project text dataset is used which has the features of the “Title of the article”, “Author of the article” and the “article text”, and “Id”. The dataset has two classes true and false. On the start of the implementation, the dataset is uploaded on the IDE (Jupiter Notebook) and the dataset has 20800 training examples and 5 features as shown in below figures is also given below:



	B	C	D
1	title	author	text
2	Specter of Trump Loosens Tongues, if Not Purse Strings, in Silicon Valley - The New York Time	David Streitfeld	PALO ALTO, Calif. â€” After years of scorning the political process, Silicon Vall
3	Russian warships ready to strike terrorists near Aleppo	nan	Russian warships ready to strike terrorists near Aleppo 08.11.2016 | Source: So
4	#NoDAPL: Native American Leaders Vow to Stay All Winter, File Lawsuit Against Police	Common Dreams	Videos
5	Tim Tebow Will Attempt Another Comeback, This Time in Baseball - The New York Times	Daniel Victor	If at first you donâ€™t succeed, try a different sport. Tim Tebow, who was a He
6	Keiser Report: Meme Wars (E995)	Truth Broadcast Network	42 mins ago 1 Views 0 Comments 0 Likes 'For the first time in history, weâ€™re
7	Trump is USA's antique hero. Clinton will be next president	nan	Trump is USA's antique hero. Clinton will be next president 08.11.2016 | Sourc
8	Pelosi Calls for FBI Investigation to Find Out â€™What the Russians Have on Donald Trumpâ€	Pam Key	Sunday on NBCâ€™s â€œMeet the Press,â€ House Minority Leader Rep. Nancy
9	Weekly Featured Profile â€“ Randy Shannon	Trevor Loudon	You are
10	Urban Population Booms Will Make Climate Change Worse	nan	Urban
11		cognitive dissident	don't we have the receipt?
12	184 U.S. generals and admirals endorse Trump for Commander-In-Chief	Dr. Eowyn	Have you
13	â€œWorking Class Heroâ€ by John Brennon	Doug Diamond	Source: CNBC, article by Robert Ferris Arctic sea ice is melting at a rate far faste
14	The Rise of Mandatory Vaccinations Means the End of Medical Freedom	Shaun Bradley	Written by Shaun Bradley Mandatory vaccinations are about to open up a new
15	Communists Terrorize Small Business	Steve Watson	Store Communists Terrorize Small Business The owner of the Blue Cat Cafe is t
16	Computer Programmer Comes Forward, Admits To Being Paid To Rig Voting Booths! TRUMP W	Usa News Flash	
17	Thieves Take a Chunk of Change, All 221 Pounds of It, From a Berlin Museum - The New York T	Melissa Eddy	BERLIN â€” You could never palm it, flip it or plunk it into a vending machine.
18	New England Patriotsâ€™ Owner, Still Sore at N.F.L., Has Payback in Sight - The New York Tim	Ken Belson and Ben Shpigel	FOXBOROUGH, Mass. â€” The N. F. L. likes portraying itself as one big family c
19	College Republicans, YAF Sue Berkeley over Ann Coulter Event - Breitbart	Tom Ciccotta	The Berkeley College Republicans and the Young Americaâ€™s Foundation hav
20	Trump Melts Down And Accuses The US Postal Service Of Stealing The Election For Clinton	Jason Easley	Trump
21	Visiting Madagascar? Leave Red Swimsuits (and Lemur Recipes) at Home - The New York Time	Bryant Rousseau	If you visit a certain beach in northeastern Madagascar, donâ€™t wear red and
22	Reeseâ€™s Peanut Butter Cups â€“ Cheap and Full of Toxic Chemicals	REALdeal	by ANYA
23	President Obama and President-Elect Donald Trump Meet at White House	REALdeal	President Obama and President-Elect Donald Trump Meet at White House: Sha
24		Dale Johnson	VERSE 9.

Figure 4. 2 Dataset



	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Figure 4. 3 Dataset

Firstly, the dataset is converted into numerical form so that we apply models on it.

4.5.1. Dataset Visualization

The dataset has 5 features and 2 classes and 20800 training examples. The dataset is well balanced between two classes of true and false and these two classes are named 0 and 1 as mentioned in the below figures 4.4.

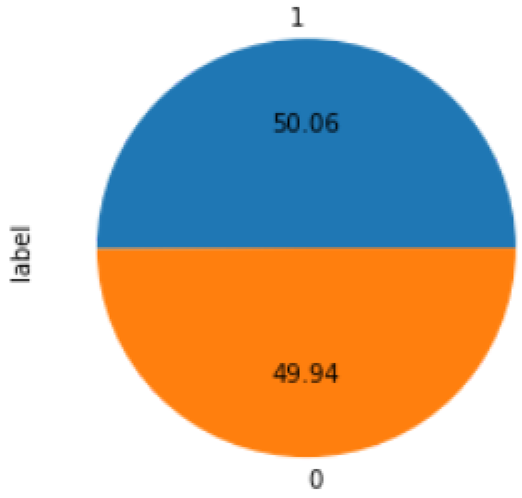
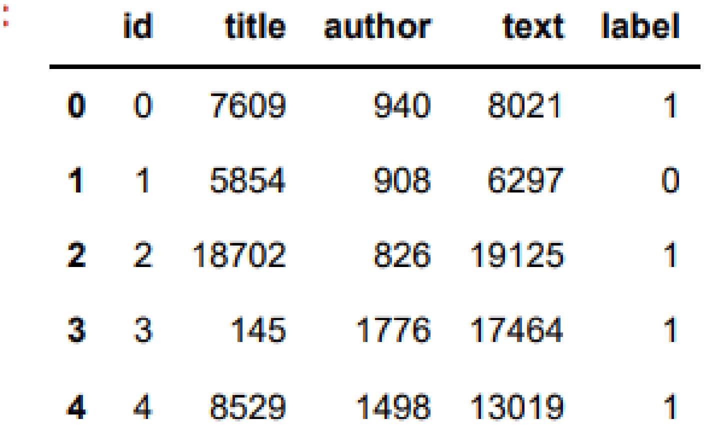


Figure 4. 4 Pie plot

In the above pie plot the blue color shows the class 1 labels which are true classes and the orange color shows the class 0 labels which are false classes. The class 1 has 50% of data examples and 0 has 49% of data examples and which shows that the dataset is well balanced between these two classes.

4.5.2. Dataset Specification

Now on the next step after uploading the dataset will be specified and label the dataset into “Title”, “Author” and “text” and will be converted into numerical form to perform preprocessing on it as mentioned in the figure 4.5 and then empty sets, cells will be removed from the dataset after which the number of rows decreases from 20800 to 18285.

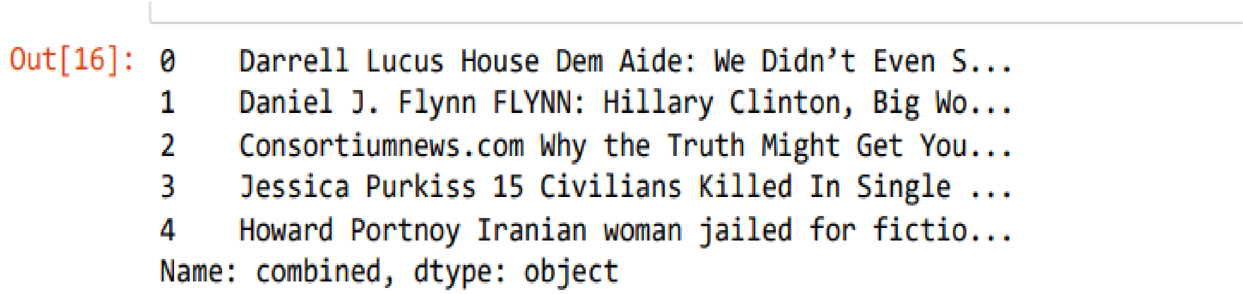


	id	title	author	text	label
0	0	7609	940	8021	1
1	1	5854	908	6297	0
2	2	18702	826	19125	1
3	3	145	1776	17464	1
4	4	8529	1498	13019	1

Figure 4. 5 Numerical Conversion

4.5.3. Features Combination

Now in the next step after specifying the dataset and removing the empty spaces and cells, in the dataset three features will be combined into one column for the purpose of applying pre-processing and below in the figure shows some combined dataset before preprocessing.



```
Out[16]: 0    Darrell Lucus House Dem Aide: We Didn’t Even S...
         1    Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
         2    Consortiumnews.com Why the Truth Might Get You...
         3    Jessica Purkiss 15 Civilians Killed In Single ...
         4    Howard Portnoy Iranian woman jailed for fictio...
         Name: combined, dtype: object
```

Figure 4. 6 Features Combinations

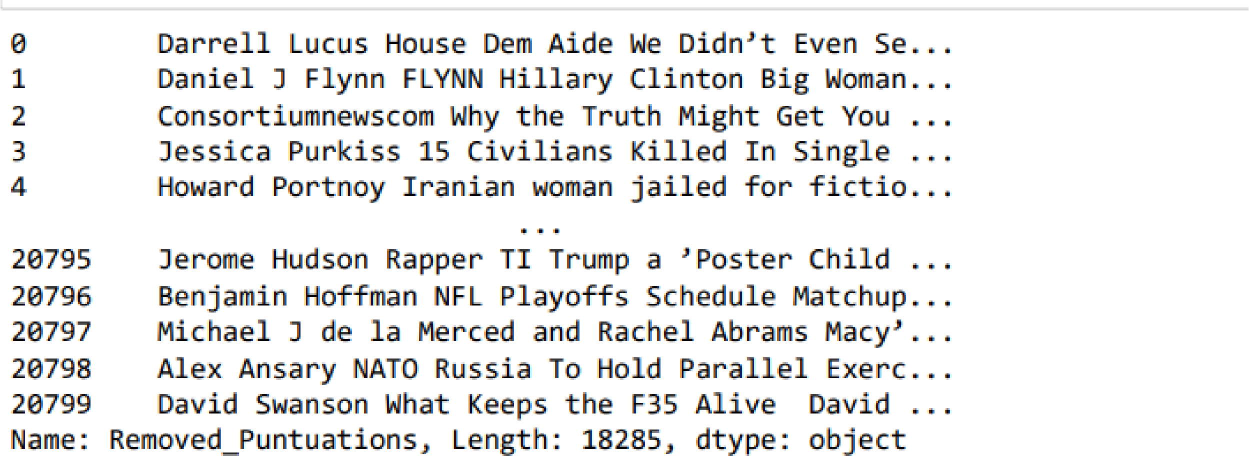
The Preprocessing step are now started and Natural Language Processing will be used for this purpose.

4.6. Natural Language Processing

Natural language processing (NLP) refers to the branch of branch of Artificial Intelligence (AI) concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP combines computational linguistics rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment. Following are some steps applied on the dataset to preprocess it for applying the Machine learning models.

4.6.1. Punctuation Removal

The first step of NLP is punctuation removal from the dataset, in which all kinds of punctuations symbols are removed for example ~`! @#$%^&*() _+= [];’’”” \/.,>< etc. as mentioned in the figure 4.7 below:

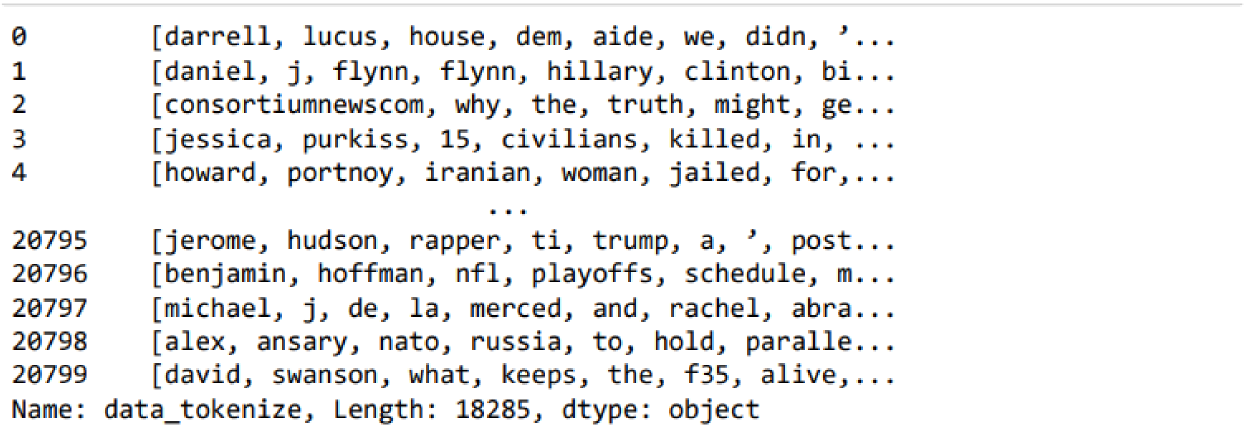


```
0        Darrell Lucus House Dem Aide We Didn’t Even Se...
1        Daniel J Flynn FLYNN Hillary Clinton Big Woman...
2        Consortiumnewscom Why the Truth Might Get You ...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                                  ...
20795    Jerome Hudson Rapper TI Trump a ’Poster Child ...
20796    Benjamin Hoffman NFL Playoffs Schedule Matchup...
20797    Michael J de la Merced and Rachel Abrams Macy’...
20798    Alex Ansary NATO Russia To Hold Parallel Exerc...
20799    David Swanson What Keeps the F35 Alive  David ...
Name: Removed_Puntuations, Length: 18285, dtype: object
```

Figure 4. 7 Punctuation Removal

4.6.2. Dataset Tokenization

After removing punctuation, the next step is tokenization in which the words in the dataset will be separated from sentences. The figure 4.8 below shows the dataset tokenization:



```
0        [darrell, lucus, house, dem, aide, we, didn, ’...
1        [daniel, j, flynn, flynn, hillary, clinton, bi...
2        [consortiumnewscom, why, the, truth, might, ge...
3        [jessica, purkiss, 15, civilians, killed, in, ...
4        [howard, portnoy, iranian, woman, jailed, for,...
                               ...
20795    [jerome, hudson, rapper, ti, trump, a, ’, post...
20796    [benjamin, hoffman, nfl, playoffs, schedule, m...
20797    [michael, j, de, la, merced, and, rachel, abra...
20798    [alex, ansary, nato, russia, to, hold, paralle...
20799    [david, swanson, what, keeps, the, f35, alive,...
Name: data_tokenize, Length: 18285, dtype: object
```

Figure 4. 8 Dataset Tokenization

4.6.3. Stop-Words Removal

Now after applying Tokenization the next step is Stop words removal in which the stop words will be removed from the dataset for example, “the”, “is” to, am “and” etc. will be removed from the dataset. The figure 4.9 below shows stop words removal from the dataset:

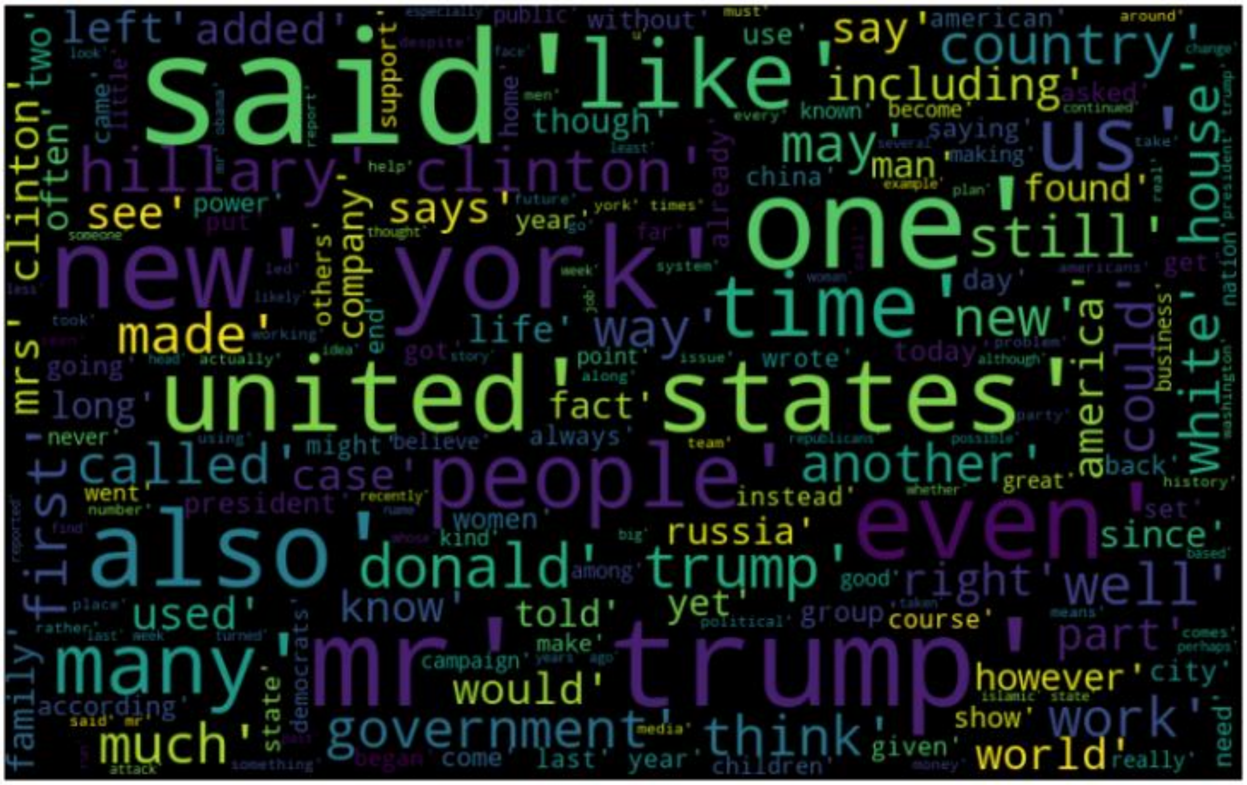
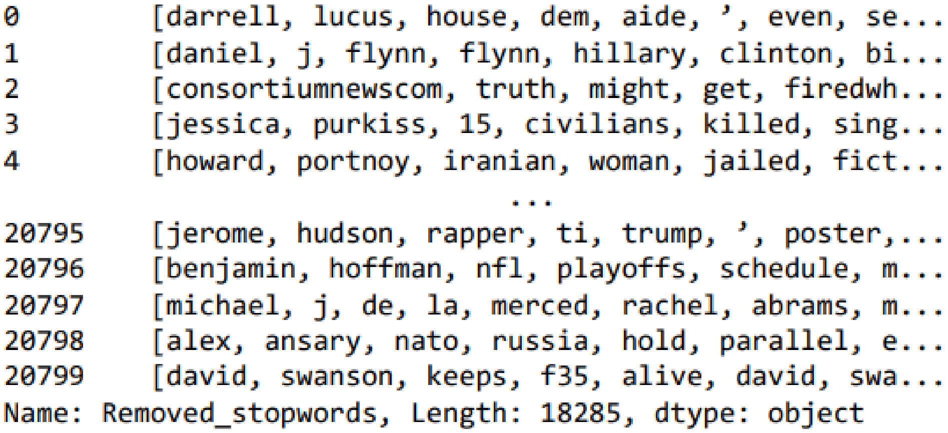


Figure 4. 9 Stop-Words

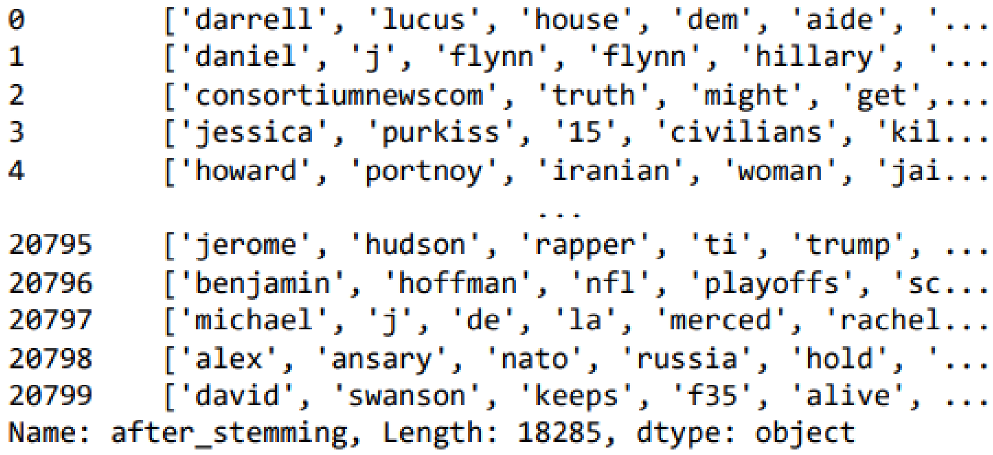


```
0        [darrell, lucus, house, dem, aide, ’, even, se...
1        [daniel, j, flynn, flynn, hillary, clinton, bi...
2        [consortiumnewscom, truth, might, get, firedwh...
3        [jessica, purkiss, 15, civilians, killed, sing...
4        [howard, portnoy, iranian, woman, jailed, fict...
                                ...
20795    [jerome, hudson, rapper, ti, trump, ’, poster,...
20796    [benjamin, hoffman, nfl, playoffs, schedule, m...
20797    [michael, j, de, la, merced, rachel, abrams, m...
20798    [alex, ansary, nato, russia, hold, parallel, e...
20799    [david, swanson, keeps, f35, alive, david, swa...
Name: Removed_stopwords, Length: 18285, dtype: object
```

Figure 4. 10 Stop-Words Removal

4.6.4. Stemming

Now after removing stop-words the next step is Stemming in which the dataset words will be normalize into their base-form. The figure 4.11 below shows the dataset stemming:

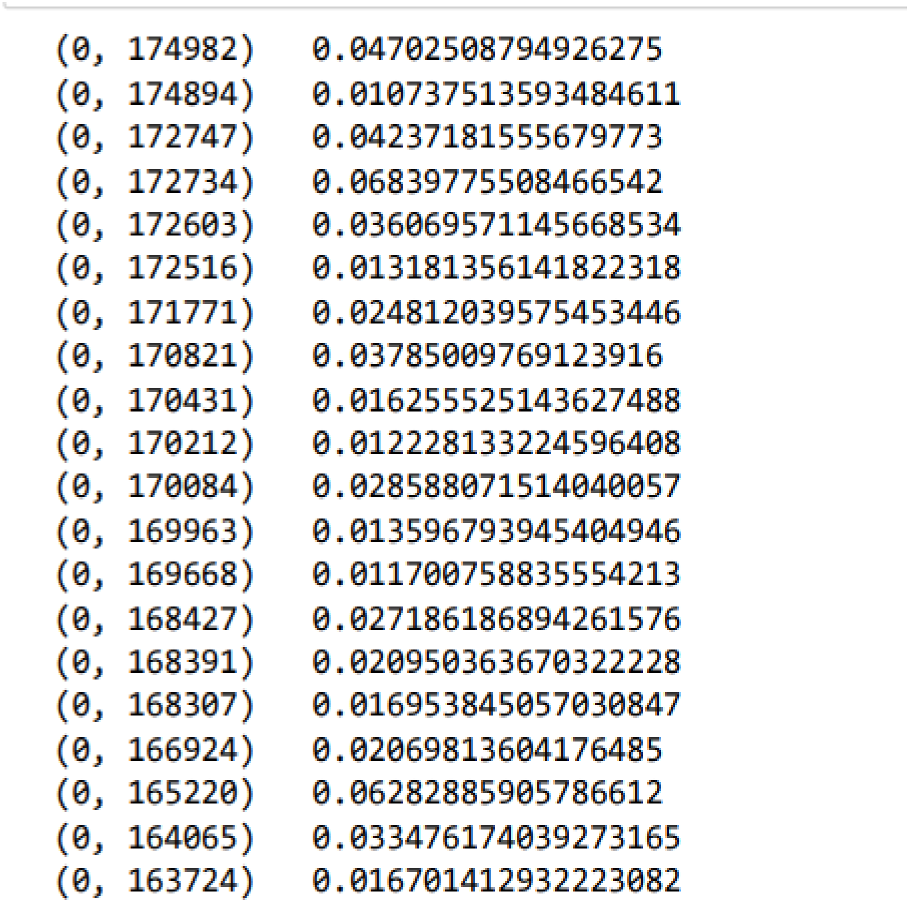


```
0        ['darrell', 'lucus', 'house', 'dem', 'aide', '...
1        ['daniel', 'j', 'flynn', 'flynn', 'hillary', '...
2        ['consortiumnewscom', 'truth', 'might', 'get',...
3        ['jessica', 'purkiss', '15', 'civilians', 'kil...
4        ['howard', 'portnoy', 'iranian', 'woman', 'jai...
                                ...
20795    ['jerome', 'hudson', 'rapper', 'ti', 'trump', ...
20796    ['benjamin', 'hoffman', 'nfl', 'playoffs', 'sc...
20797    ['michael', 'j', 'de', 'la', 'merced', 'rachel...
20798    ['alex', 'ansary', 'nato', 'russia', 'hold', '...
20799    ['david', 'swanson', 'keeps', 'f35', 'alive', ...
Name: after_stemming, Length: 18285, dtype: object
```

Figure 4. 11 Stemming

4.6.5. Vectorization

Now after applying Stemming, Vectorization will be applied in which features will be extracted and the below figure 4.12 shows the dataset vectorization:



```
  (0, 174982)	0.04702508794926275
  (0, 174894)	0.010737513593484611
  (0, 172747)	0.0423718155679773
  (0, 172734)	0.0683977550846542
  (0, 172603)	0.036069571145668534
  (0, 172516)	0.013181356141822318
  (0, 171771)	0.024812039575453446
  (0, 170821)	0.0378500976912391
  (0, 170431)	0.016255525143627488
  (0, 170212)	0.012228133224596408
  (0, 170084)	0.028588071514040057
  (0, 169963)	0.013596793945404946
  (0, 169668)	0.011700758835554213
  (0, 168427)	0.027186186894261576
  (0, 168391)	0.020950363670322228
  (0, 168307)	0.016953845057030847
  (0, 166924)	0.0206981360417648
  (0, 165220)	0.0628288590578661
  (0, 164065)	0.033476174039273165
  (0, 163724)	0.016701412932223082
```

Figure 4. 12 Vectorization

4.7. Logistic Regression

Logistic regression guesstimates the possibility of an event happening, such as Fake or real, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. So, we have applied this machine learning classifier on our dataset and it gives the 95% of accuracy.

4.8. Support Vector Machine

A support vector machine (SVM) is a supervised machine learning model used for two group classification problems. SVM categorized the data when we get the dataset labeled. We also applied SVM to check the accuracy and we got an accuracy score of 99% on testing the dataset.

4.9. Naive Bayes

Naïve Bayes is a supervised classification algorithm that is used for binary and multi-class classification. It uses conditional probability to classify future objects by passing labels to training examples. We also applied this algorithm to check the accuracy and we got an accuracy score of 76% on testing the dataset.

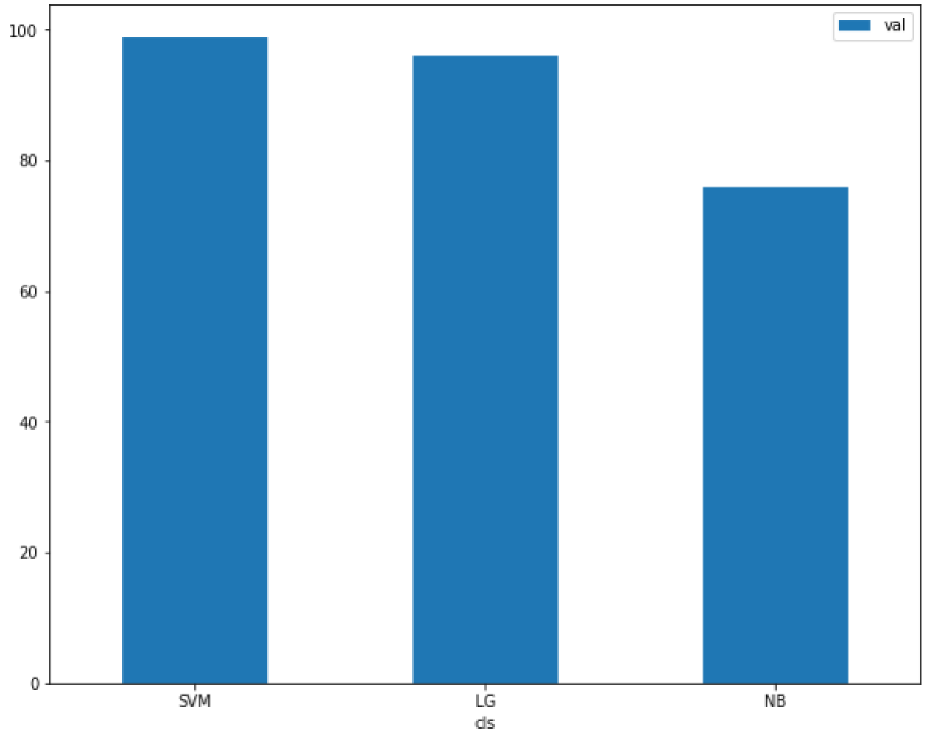


Figure 4. 13 Combine Accuracy

The above figure 21 shows the combined accuracy of all three models.

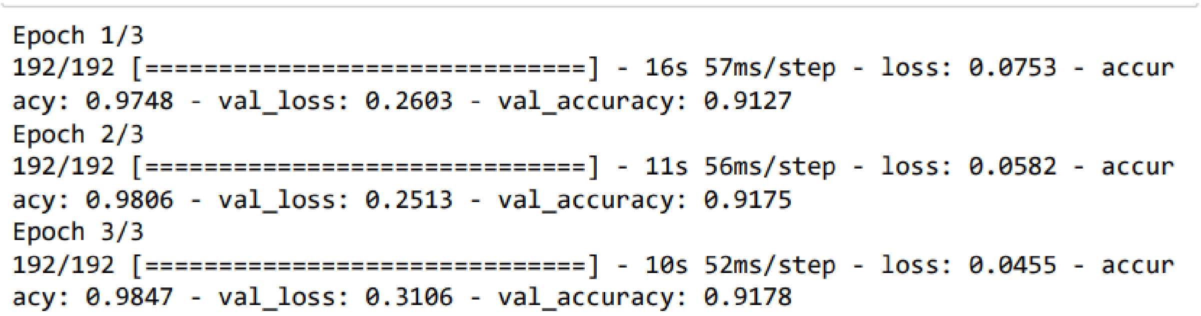
4.10. ‘Recurrent Neural Network

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (NLP), speech recognition, and image captioning; they are incorporated into popular applications such as Siri, voice search, and Google Translate. Recurrent neural networks utilize training data to learn. They are distinguished by their “memory” as they take information from prior inputs to influence the current input and output.

While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depends on the prior elements within the sequence. While future events would also help determine the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions. We also applied this algorithm to check the accuracy and we got an accuracy score of 97% on testing the dataset.

4.11. ‘Long-Term Short Memory

LSTM networks are a type of RNN that uses special units in addition to standard units. LSTM units include a ‘memory cell’ that can maintain information in memory for long periods. This memory cell lets them learn longer-term dependencies and it is a famous deep learning algorithm that is well suited for making predictions and classification with a flavor of the time. The advantage of the Long Short-Term Memory (LSTM) network over other recurrent networks is an improved method of back-propagating the error. After RNN we applied this algorithm to check the accuracy and we got a 91% accuracy score while we got an accuracy score of 98% on testing the dataset.



```
Epoch 1/3
192/192 [==============================] - 16s 57ms/step - loss: 0.0753 - accur
acy: 0.9748 - val_loss: 0.2603 - val_accuracy: 0.9127
Epoch 2/3
192/192 [==============================] - 11s 56ms/step - loss: 0.0582 - accur
acy: 0.9806 - val_loss: 0.2513 - val_accuracy: 0.9175
Epoch 3/3
192/192 [==============================] - 10s 52ms/step - loss: 0.0455 - accur
acy: 0.9847 - val_loss: 0.3106 - val_accuracy: 0.9178
```

Figure 4. 14 LSM

Chapter 5 Testing and Analysis

5.1 Testing Techniques Employed for This Project

Failures or errors can occur in a system. To avoid this, it is vital to test the system once it has been developed and overcome any difficulties that may arise. In this system, we also ran testing to see if there are any problems. We tested the system's accuracy and precision using test scenarios.

5.2 Integration

It focuses on the system's input and output, such as characteristics gathered through detection and output. An algorithm that evaluates the system on a test dataset and provides accuracy has been added to this system.

5.3 System Result

The entire system is tested to ensure that each function operates as designed. The cases represent the test results.

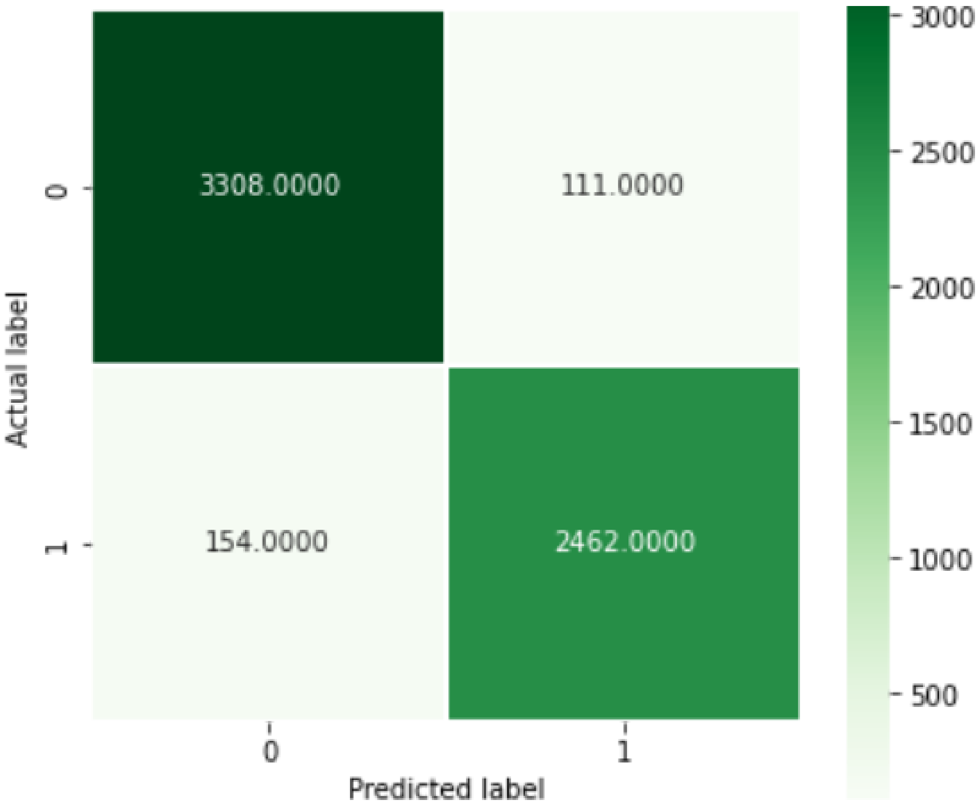


Figure 5. 1 Test Case 1

The above figure 5.1 shows the heatmap or confusion matrix of fake new detection on social media using logistic regression, in which the dark green and the light green color shows the intensity of the number of correct predictions are higher and white color shows that the intensity of the number of wrong predictions are lower.

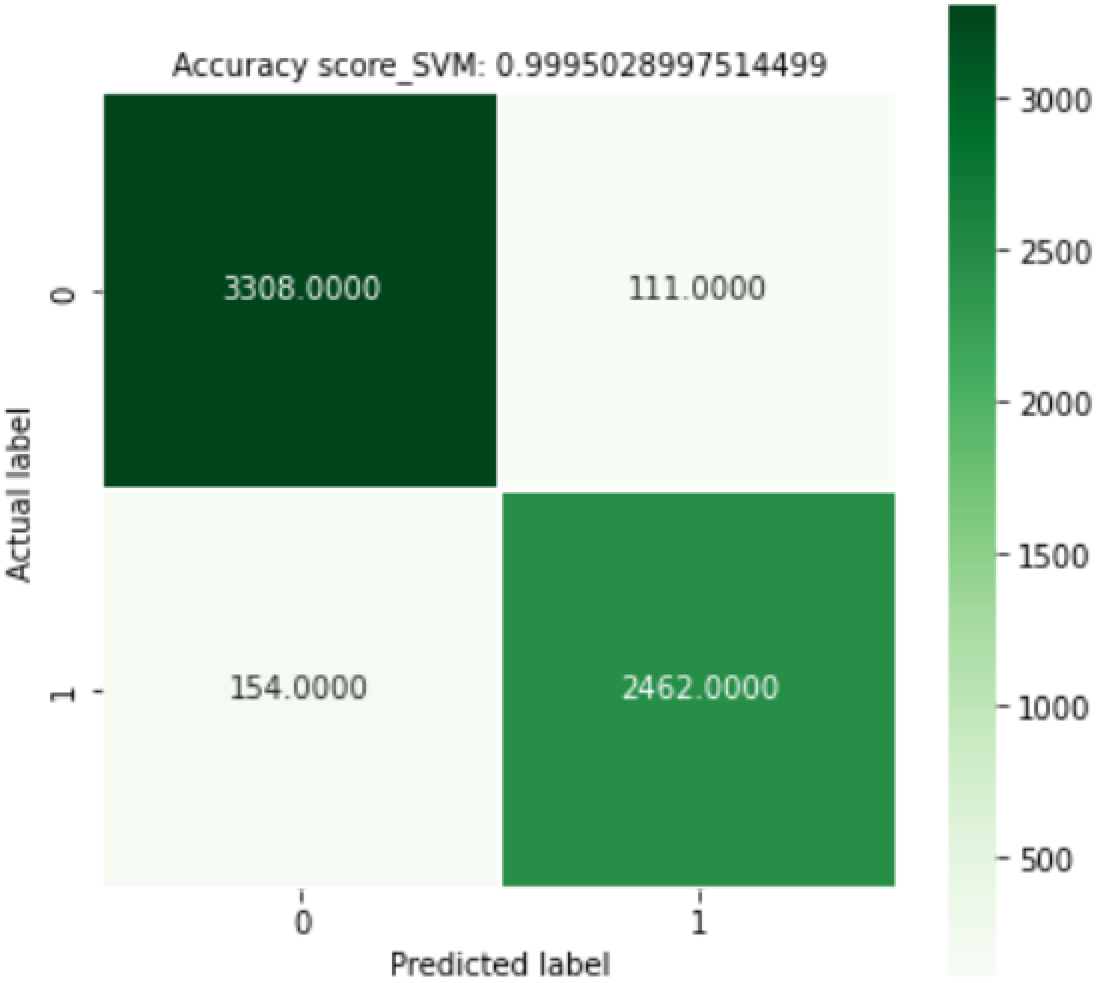


Figure 5. 2 Test Case 2

The above figure 5.2 shows the heatmap or confusion matrix of fake news detection on social media using the Support Vector machine, in which the dark green and the light green color shows the intensity of the number of correct predictions are higher and the white color shows that the intensity of the number of wrong predictions is lower.

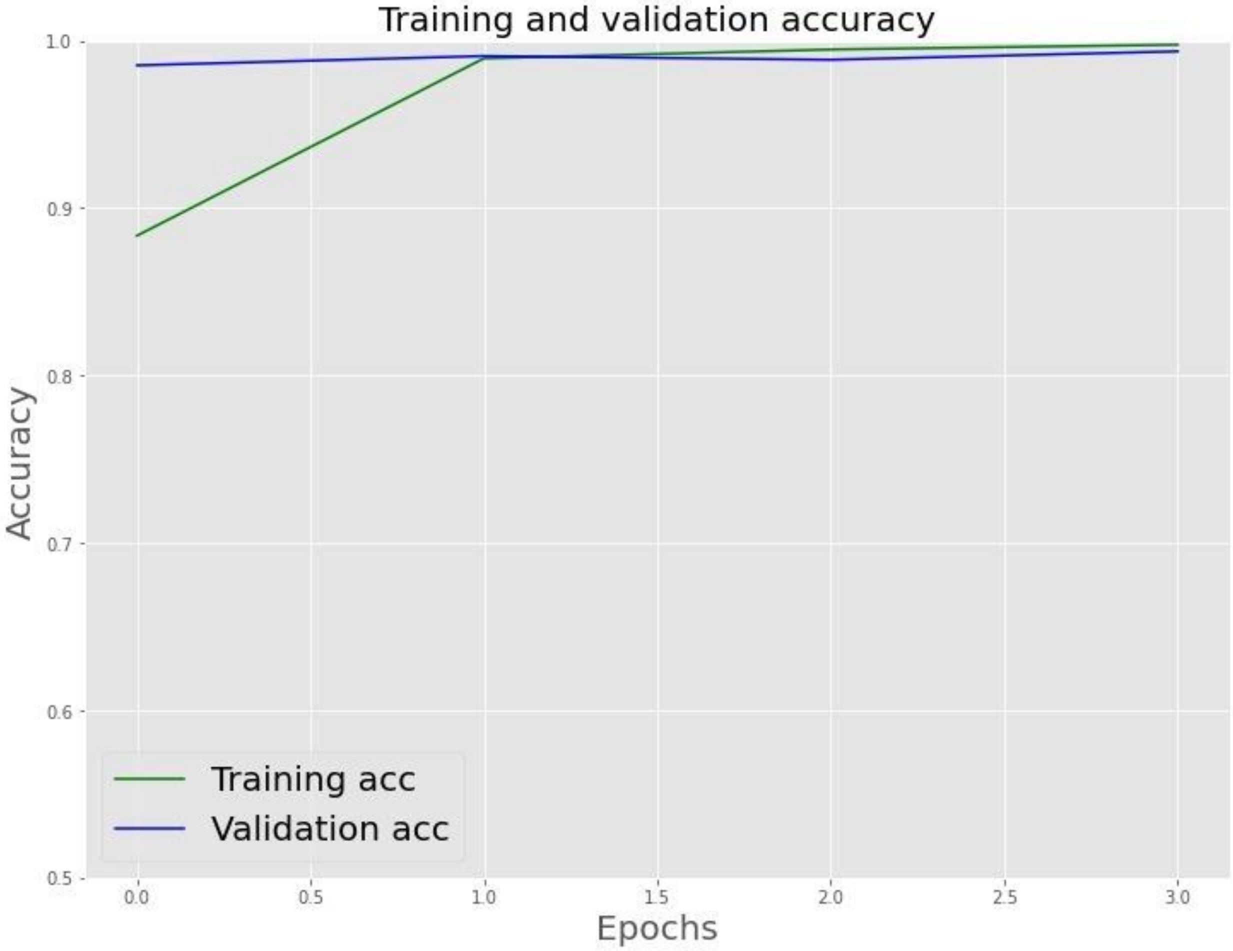


Figure 5. 3 RNN

The above figure 5.3 represents the difference in the training accuracy and validation accuracy of the Recurrent Neural Network model. In the above graph, the green line shows training accuracy and the blue line represents the validation accuracy, which shows that at the start of the epochs the training accuracy is low but when the number of epochs increases the training accuracy also increases but on the other hand the validation accuracy is high on the start and so on.

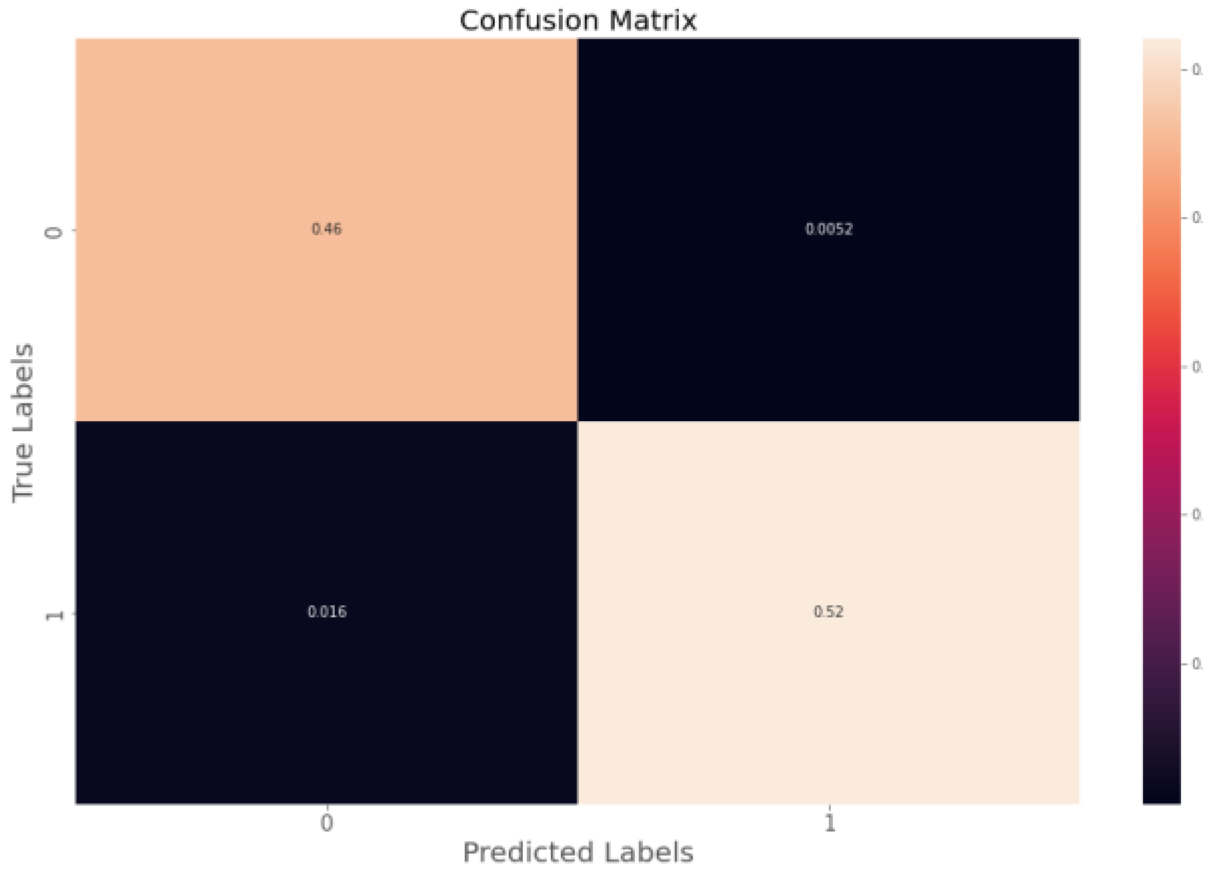


Figure 5. 4 RNN Confusion Matrix

The above figure 5.4 shows the heatmap or confusion matrix of fake news detection on social media using the RNN, in which the dark and the light-dark color show the intensity of the number of correct predictions is higher and the white color shows that the intensity of the number of wrong predictions is lower.

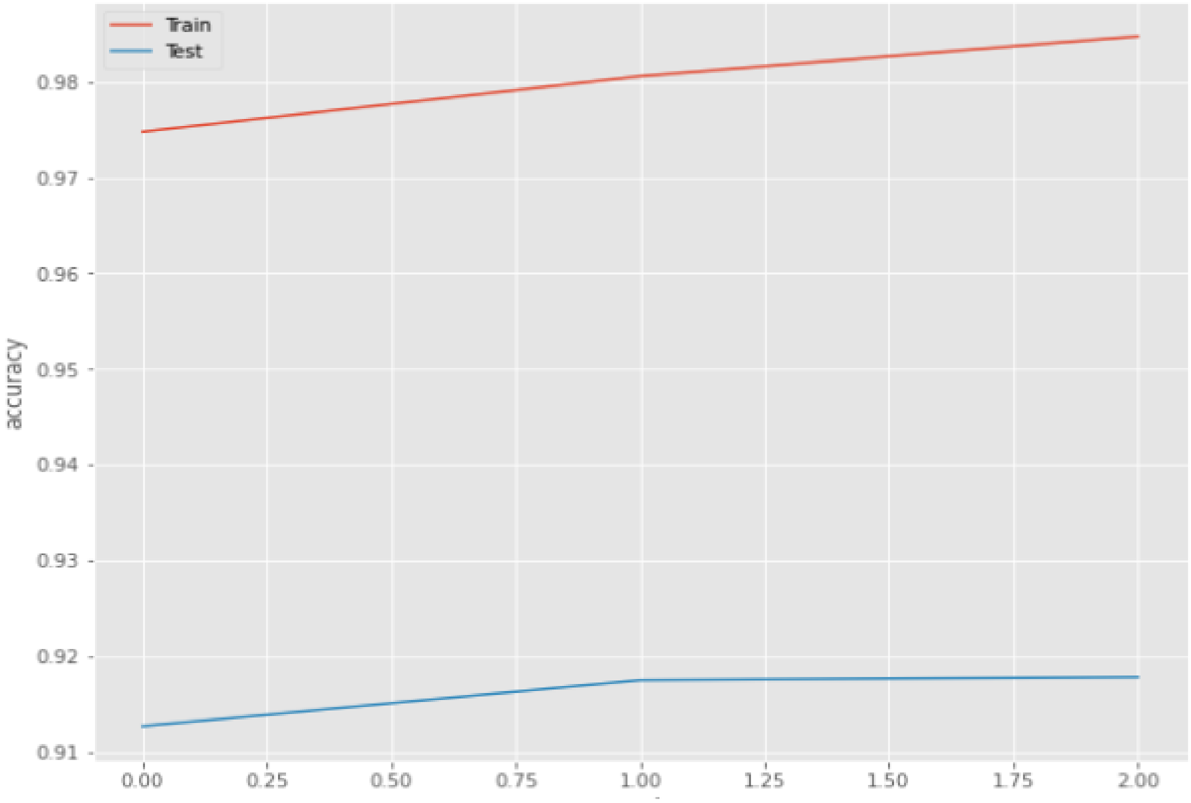


Figure 5. 5 LSTM Results

The above figure 5.5 shows train and test accuracy on the dataset using LSTM, the red line shows that the training accuracy is high as the number of epochs increases by 99% and the blue line shows that the test accuracy remains almost constant as the number of epochs increases of 92%.

5.4 Summary

We intend to research fake news detection on social media utilizing a combination of Machine learning, NLP, and Deep Learning technologies such as Logistic Regression, SVM and Naïve byes, and LSTM. Then we examined their accuracy one by one, and virtually all of the models had an accuracy. We used LSTM and RNN to enhance the accuracy of the dataset. To do this, a series of actions must be taken. This work presents those stages that are critical and used in studies on RNN and LSTM, although there may be certain difficulties when employing detection to recognize text datasets, this dataset is freely accessible. The collection comprises test of articles. After testing the functionality, we can conclude that the system provides the highest accuracy when Machine Learning models and NLP are combined and with Deep learning models utilized in different feature extraction methods.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

Finally, the algorithm is created utilizing the machine learning & deep learning model & NLP. Based on the text articles dataset, it is more effective to identify the fake or real news articles. The method is very simple to use, accurate, and user-friendly. For the user, it's a trustworthy system. The system is in good condition and satisfies all the requirements. After testing the functionality, we can conclude that the system provides the highest accuracy when Machine Learning models and NLP are combined and with Deep learning models utilized in different feature extraction methods. The overall performance is outstanding, and it may be used to detection of fake news with accuracy.

6.2 Future Work

As we all know, the machine learning & NLP area moves fast, yet a small amount of work is done in deep learning. As a result, there are several options to engage with Fake news detection. As a new era of technology approaches, computer processing power will increase, and training time will decrease. With better processing performance, our system will run more efficiently. If there is enough processing capacity, I can use a large dataset with my model.

6.3 Improvements in the System

The system is now processing limited datasets, but this may be expanded in the future to function on a larger scale. Training on a huge dataset can enhance accuracy. The method should be accurate for different types of fake new articles that are labeled as fake or real in training text articles.

References

[1] Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, Amitrajit Sarkar.," Analysis of Classifiers for Fake News Detection", Procedia Computer Science " Inc., publishers, 2019.

[2] I. Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods," 2020.

[3] M. A. Liebert, FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media, Mary Ann Liebert, Inc., publishers, 2020.

[4] Alim Al Ayub Ahmed, "Detecting Fake News using Machine Learning:," 2010.

[5] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, Yan Liu." Combating Fake News", ACM Transactions on Intelligent Systems and Technology", 2019.