Fake News Detection Using ML

project Report submitted in partial
fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

Submitted by

Niket Kumar
1713120006/17SCSE120003
IN
Computer science & Engineering with specialization of Big Data

Navneet Himanshu
1713106004/17SCSE1122019
IN
Computer science & Engineering with specialization of Data analytics

School of computing science & Engineering

Under the supervision of
Dr. Jayakumar V
Assistant professor

GALGOTIAS
UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

May/June 2021

GALGOTIAS
UNIVERSITY

School of computing science & Engineering

BONAFIDE CERTIFICATE

Certified that this project report **Fake News Detection Using ML”** is the bonafide work of “**Niket Kumar,Navmeet Himanshu”** who carried out the project work under my supervision

	V. Jayakumar
SIGNATURE	SIGNATURE
<<Name>>	Dr. Jayakumar . V
Dean of School	SUPERVISOR Assistant Professor
<<School Name>>	School of Computing Science and Engineering

Approval Sheet

This thesis/dissertation/report entitled (Fake News Detection Using ML) by (Niket Kumar,Navneet Himanshu) is approved for the degree of B.Tech(Hons)

Examiners

____

____

V Jaya

Supervisor (s)

____

____

____

Chairman

____

Date:_________

Place:_________

Statement of Project Report Preparation

1. Thesis title: Fake News Detection Using ML

2. Degree for which the report is submitted: B.Tech(Hons) in CSE with specialization in Big Data

3 Project Supervisor was referred to for preparing the report.

4. Specifications regarding thesis format have been closely followed.

5. The contents of the thesis have been organized based on the guidelines.

6. The report has been prepared without resorting to plagiarism.

7. All sources used have been cited appropriately.

8 The report has not been submitted elsewhere for a degree.

(Signature of the student)

Niket Kumar

Name:Niket Kumar

Roll No:1713120006

Statement of Preparation:

I have prepared The project: Fake News Detection using Ml” on jupyter Notebook. The aim of this project is to detect the fake news present in the article. For this I have studied ideas from many sites which makes easy for implementation.

Abstract

This Project “Fake News Detection” works on the applications of Natural Language Processing(NLP) techniques that recognizes the 'fake news', that is deceptive news stories which comes from the unidentified sources. During this systematic review, the factors that results in the spreading of fake news and information have been provided. In this report, the identification of the basic cause which results in the spreading of fake news are performed which may result in the break of fake news among public domain. In order to conquer the social media platform from the rapid spread of fake news, firstly we should know the reason and intention behind the spreading of fake news. Therefore, this review takes associate in early initiative to find the major reason which lead to the expansion of fake news among public domain. The main aim of this review is to find out with what intention and why people unknowingly share information which may be false and to presumably facilitate in detection of fake news before it spreads. There the model should be build which support a count vectorizer or a (TFIDF) Term Frequency Inverse Document Frequency matrix, will solely get you up to now. However sometimes these following models did not consider the important qualities like ordering of word and context. It may be possible that 2 articles whose word counts may be similar are totally alter in their meanings.

List of Figures

Table of Contents

CHAPTER 1

1. Introduction

The emergence in the swift increase of internet users and also the fast acquisition of social media platforms such as twitter and Facebook sealed the method for circulation of information that has never been witnessed within the human history earlier. Fake news refers to false or misleading information content whose source cannot be verified. Besides different instances, news retailers gained from the widespread use of different media network platforms by delivering updated news to their subscribers through apps, and different digital platforms they can be either Facebook or website or WhatsApp or twitter, blogs, social media feeds, and other digital media platforms. Social media platforms such as twitter, Facebook, Instagram, WhatsApp, etc are considered so much influential when it comes to news feed. This became quiet for customers to accumulate the most recent news very quickly. These social network platforms are gaining such huge popularity because in today's time they provide an edge to their users that they can express their feelings as well as discuss together and represent their thought in front of society, they ca share their opinion over the topics like health, poverty, education and literacy. But these social media platforms such as Facebook, WhatsApp, twitter can also be used in negative manner where they can used to spread fake news across the society, negativity among the youth which can lead to riots as we have seen many cases and these cases are increasing every year exponentially. These fake news on one place can cause mass destruction but for somebody present in between of us can earn money by selling such news.

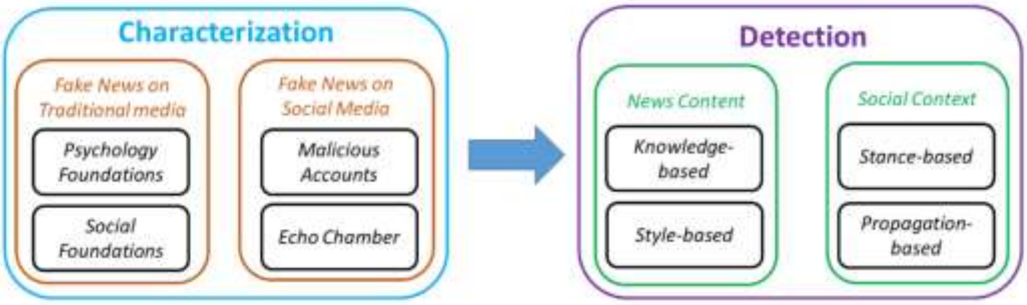


Figure 1.1: Fake news on social media: from characterization to detection.

Fake news is not only limited to small point they got so much popularity during the time of USA election 2016 as well as 2021. Huge mass of population was sharing information among each other through social media mainly facebook without knowing the facts which lead to serious issues. These fake news articles were not just covering politics but much more. They have alternative domains too such as heals, sports, science as well as lifestyle too. Financial market is the worst affected area from fake news and articles, Over here a small fake news can be disastrous which can lead to a halt on the market which will be loss making. This world is formed by the data or information which we can digest. There is a proof that customers act angrily over the news which later found out to be fake or false or incorrect. Recent fake news which was spreading was related to the topic Novel coronavirus. There were so many misinformations regaring the behaviour of virus, it's spread as well as many aspects which were fake but provided to very large chunk of population.

1.1 Feature Extraction

1.1.1 News Content Features

Now that fake news has been defifined and the target has been set, it is needed to analyse what features can be used in order to classify fake news. Starting by looking at news content, it can be seen that it is made of four principal raw components:
• **Source**: Where does the news come from, who wrote it, is this source reliable or not.
• **Headline**: Short summary of the news content that try to attract the reader.
• **Body Text**: The actual text content of the news.
• **Image/Video**: Usualy, textual information is agremented with visual information such as images, videos or audio.

Features will be extracted from these four basic components, with the mains features being linguistic-based and visual-based. As explained before, fake news is used to inflfluence the consumer, and in order to do that, they often use a specifific language in order to attract the readers. On the other hand, non-fake news will mostly stick to a diffferent language register, being more formal. This is linguistic-based features, to which can be added lexical features such as the total number of words, frequency of large words or unique words.

The second features that need to be taken into account are visual features. Indeed, modifified images are often used to add more weight to the textual information. For example, the **Figure 1.2** is supposed to show the progress of deforestation, but the two images are actually from the same original one, and in addition the WWF logo makes it look like to be from a trusted source.

1.1.2 Social Context Features

In the context of news sharing on social media, multiple aspect can be taken into account, as user aspect, post aspect and group aspect. For instance, it is possible to analyse the behaviour of specifific users and use their metadata in order to fifind if a user is at risk of trusting or sharing false information. For instance, this metadata can be its centre of interest, its number of followers, or anything that relates to it

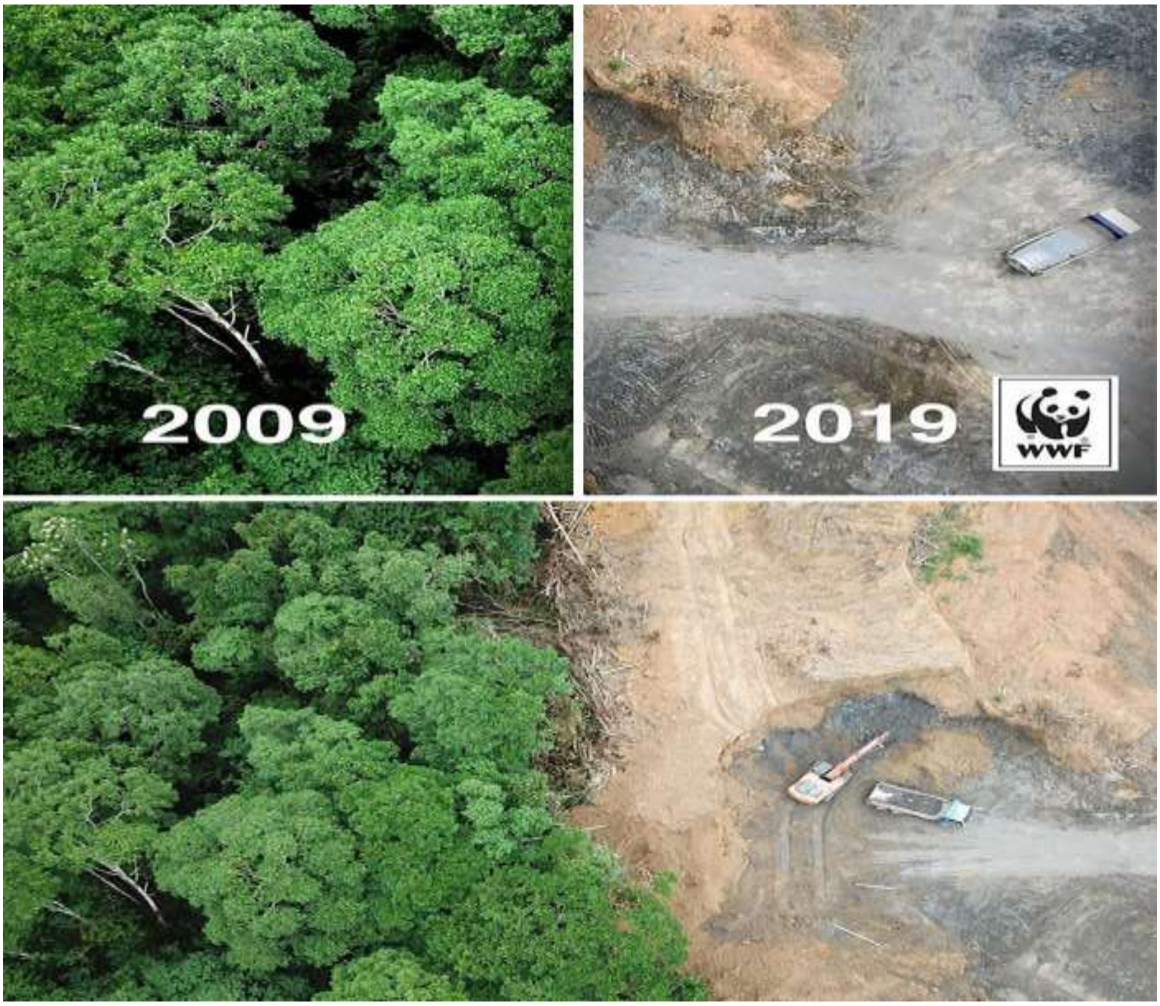


Figure 1.2: The two images provided to show deforestation between two dates are from the same image taken at the same time.

1.2 Contribution

In the compilation of fake news, we need to consider multiple instances where both supervised and unsupervised machine learning is used for the classification of texts. Though mostly the literature emphasizes upon certain specific domains like political domains. On getting the result we can see our algorithm works better for a particular type of domain and gives optimum results but when it focuses on different type of domains, we can observe it does not give optimum results. Since the articles that are from different domains have their own way of representation of textual structure it causes a problem to train a general algorithm which will give the best results when dealt with different sorts of domains. So, in this paper we are proposing the solution regarding detection of fake news problem where we are going to use machine learning ensemble method. On Studying we have come across towards different properties which we are going to use in order to distinguish between fake news and real news. With help of these properties, we are going to train a mixture of different Machine Learning Algorithms through varied methods This has been proved to be helpful when we have to choose a large choice of applications and by implementing famous techniques or methods like bagging and boosting because these learning models can reduce or decrease the error rate to a minimum level. These techniques or methods proofs to be easier and more efficient during the training of different machine learning algorithms.

Literature Review

❖ Fake News Impact:

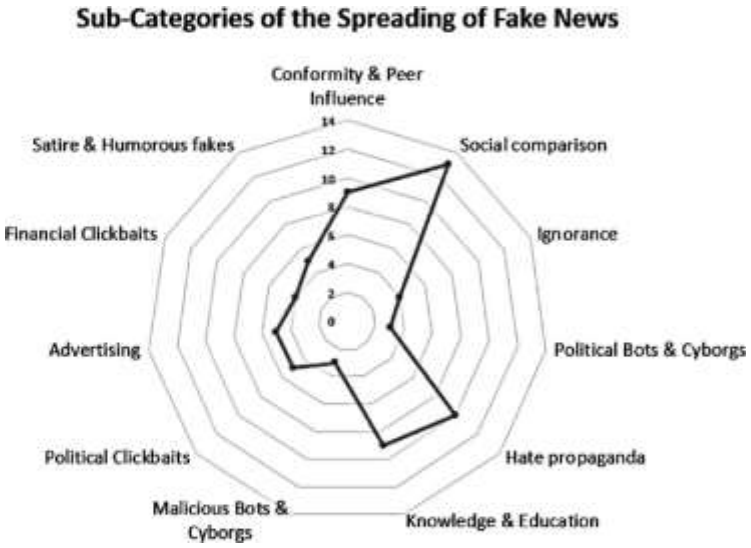
In today's world, internet is driven by news and advertisement. Websites which contains hot news and sensational headlines results in advertising helps in capitalizing the high traffic to the site. It has been seen that fake news websites makers made money with the help of automated advertising that rewards them high traffic to their websites. The continuous spreading of information causes stress and confusion among the public or geenral citizens. fake news that deliberately created to cause damage and to mislead the general public is called or known as digital misinformation. Misinformation has a very high potential to cause problems, among minutes, for variant of people. Misinformation has been famed to cause disputes, to disrupt elections, to produce unease, and the moost important hostility among the general public.

❖ Fake News influenced by social platforms:

In today's world or society, the internet became a significant or very important part of our society or in simple words of our daily lives. Anyone can easily receive the modified form of Information so easily through Social Media such as Whatsapp, Facebook etc. It was also reported in 2016 that Whatsapp and Facebook is the a very big or giant social media platform, which consists more than 20 lakh users around the world. In spreading of the fake news it has been seen that the role played by Facebook & whatsapp has the greatest impact as compared to all the other social media platforms such as twitter, orkut,etc. It had also been reported that approximately half of globe users get their news from Facebook and 22% from whatsapp. 25% of Facebook users have indicated that they have shared the false information, in many forms which also include accidental form. The spread of such news is mainly with the help of social media platforms and it's happening at a very very fast speed or pace.

❖ Analysis of Findings

In this paper, possible reasons for and factors contributing to the sharing and spreading of false information are discussed. The reasons are categorized under various factors highlighted in the journal articles used to answer the research question. These factors include: social factors, cognitive factors, political factors, financial factors and malicious factors.



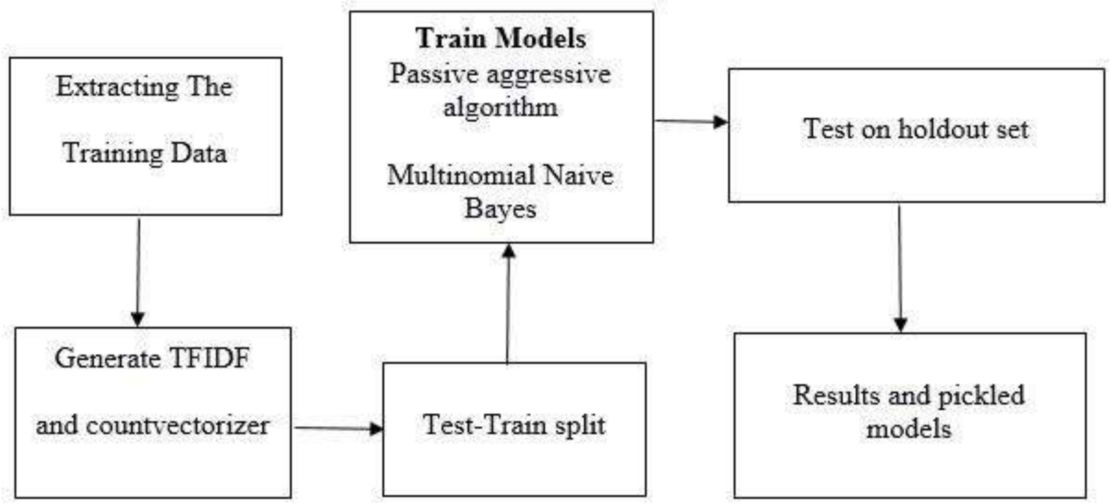
PROBLEM FORMULATION:

In this paper, we are required to study and research about the fake news detection which includes the creators , articles, and problem subjects in different or various online social media forms and platforms. Based on various forms of this diversified data sources, therefore the article subject and various authorship & the most important relationships among them, we are required to aim for different different fake news which are generating from social networks continuously. we are required to find out a way to identify the fake news for getting good quality content, where only the important and genuine news can exist whereas fake ones either identified or be removed or scraped from social media platform. Dealing with fake news detection isn't a very easy task or a cake walk because of the following reasons

Use of Textual data: For the articles related to news, creators of those articles and subjects, a collection or group of the information in terms of text regarding their created contents such as news, advertisement, description and profile will be collected from social media such as facebook, whatsapp, twitter and various other platforms. To perform all these actions we require a model with best feature extraction and learning ability qualities.

Fusion of Heterogeneous Information: The labels of credibility of creators and news articles have a very very strong connections, which we can indicate by or with the help of the article subject & authorship relationships among them. An efficient and very effective way of such correlations in our model learning will be very helpful for so much precise results regarding fake news.

Problem Formulation: Given a News-HSN G = (V, E), the fake news detection problem aims at learning an inference function $f : U \cup N \cup S \rightarrow Y$ to predict the credibility labels of news articles in set N , news creators in set Uand news subjects in set S. In learning function f, various kinds of heterogeneous information in network G should be effectively incorporated, including both the textual content/profile/description information as well as the connections among them.


Extracting The Training Data
Train Models
Passive aggressive algorithm
Multinomial Naive Bayes
Test on holdout set
Generate TFIDF and countvectorizer
Test-Train split
Results and pickled models

CHAPTER 2

2. Software Requirements

- **Jupyter Notebook:**The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text

- **NUMPY:** It is a python library used for performing high level mathematical numerical operations and it also supports multidimensional arrays and matrices. It was developed by Jim Hugunin.
- **MATPLOTLIB:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

- **PANDAS :** It is a python library used for data analysis. It gives greatly optimized performance and it is written in c or python. It is used in academics and commercial domain.

❖ Hardware Interfaces

1. **Processor :** Intel CORE i5 processor with minimum 2.9 GHz speed.
2. **RAM :** Minimum 4 GB.
3. **Hard Disk :** Minimum 500 GB

Feasibility Analysis

Model Performance: Our designed model is based on classification and is aimed to produce efficient results.

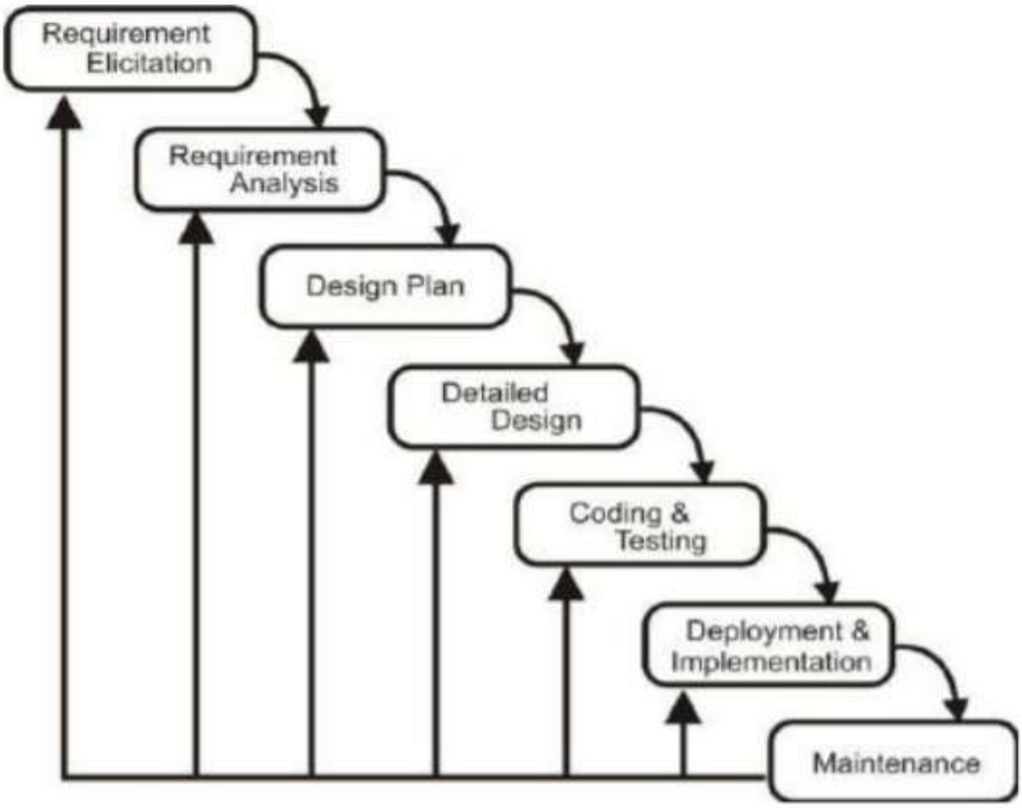
- **Technological considerations:** The analysis will be performed on a large set of data and from that only reliable sources are taken into consideration.
- **Financial feasibility:** The model is less expensive as we gather information from government sites which are free to access and we get the structured data from kaggle also. A large staff is also not required as the software only requires basic concepts to work on.
- **Resource feasibility:** The model is primarily depended on large data sets. So, having large resources will maximize the result more effective.

CHAPTER 3

1. Analysis:

The steps that we followed while developing this project are-:

- Analysis of the problem statement.
- Gathering of the requirement specification
- Analyzation of the feasibility of the project
- Development of a general layout.
- Going through the journals regarding the previous related works on this particular field.
- Choosing the method for developing the algorithm.
- Analyzing the various pros and cons.
- Starting the development of the project
- Installation of software like Jupyter Notebook
- Analyzation of algorithm by guide.
- Coding as per the developed algorithm.



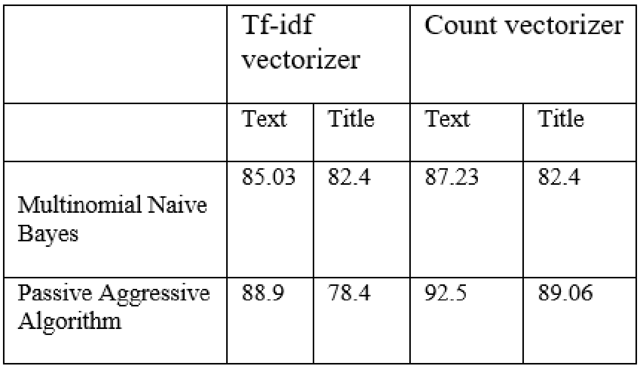
CHAPTER 4

METHODOLOGY

This paper explains the system which is developed in three parts. The first part is static which works on machine learning classifier. We studied and trained the model with different classifiers and chose the best classifier for final execution. The second part is dynamic which takes the keyword/text from user and searches online for the truth probability of the news. The third part provides the authenticity of the URL input by user.In this paper, we have used Python and its Sci-kit libraries . Python has a huge set of libraries and extensions, which can be easily used in Machine Learning. Sci-Kit Learn library is the best source for machine learning algorithms where nearly all types of machine learning algorithms are readily available for Python, thus easy and quick evaluation of ML algorithms is possible.

Proposed system Merits(Advantages):

Accuracy achieved by the different methods are:



	Tf-idf vectorizer		Count vectorizer	
	Text	Title	Text	Title
Multinomial Naive Bayes	85.03	82.4	87.23	82.4
Passive Aggressive Algorithm	88.9	78.4	92.5	89.06

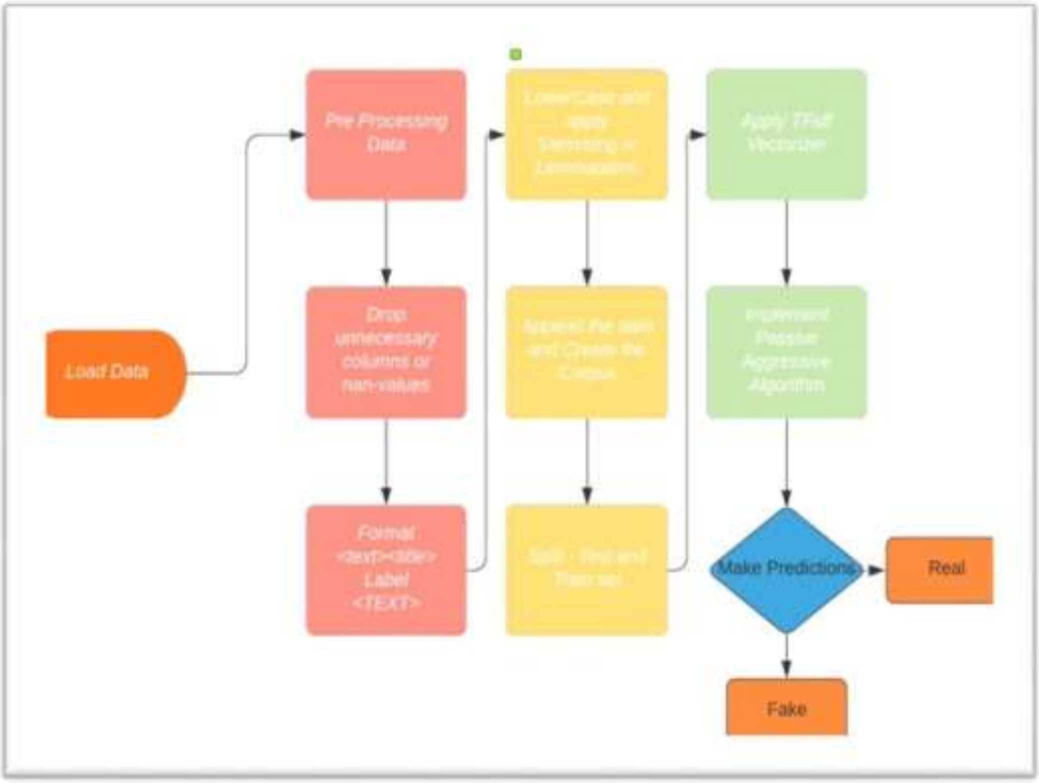
From the above accuracy table it is found that choosing count vectorizer and implementing the Passive Aggressive Algorithm on the Text data results in more accuracy.

Accuracy:
Accuracy is simply the metric which is represnting, proportions of properly and precisely predicted observations or results. To calculate the accuracy of the model's performance, the below equation is often used or mostly used:Most of the time, high accuracy is the good and more efficient and effective model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP=True Positive
FP=False Positive
TN=True Negative
FN=False Negative

Architecture Diagram:


Load Data
Pre Processing Data
Drop unnecessary columns or nan-values
Format <text><title> Label <TEXT>
LowerCase and apply Stemming or Lemmatization
Append the data and Create the Corpus
Split - Test and Train set
Apply TFidf Vectorizer
Implement Passive Aggressive Algorithm
Make Predictions
Real
Fake

CHAPTER 5

Implementation & Testing

Pre processing of the given data:

In pre-processing step our first step is to import libraries then load the dataset which is open source dataset and we got it from kaggle in Jupyter Notebook.From the dataset extract the dependent and independent variables .By importing the train test split the process of splitting of dataset into training and test set occurs.After performing this process with code pre-processing of data completes.

Fitting of the Decision tree algoritm to the Training set:

In this process we fit the proposed model to the training set.For implementing this firstly import the Decision Tree Classifier class from sklearn.tree library.

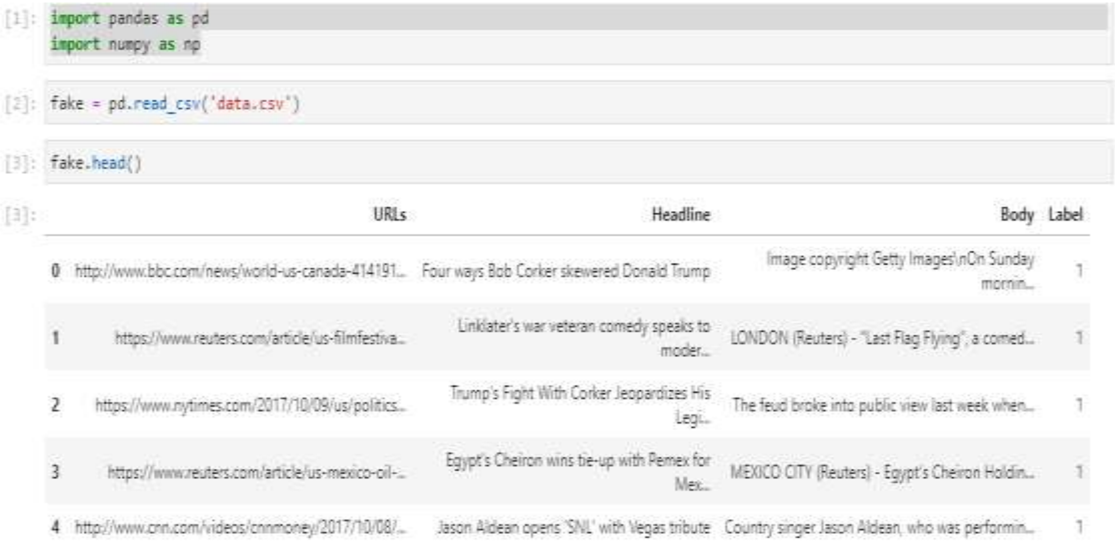
Predicting the test-set result:

In this we will try to predict the test-set result by fitting the model by creating a new prediction vector.

Confusion matrix:
After getting the output we might get some incorrect predictions so for this we have to know the number of correct and incorrect predictions.for this we need the confusion matrix and for this we import the confusion matrix from sklearn.metrics.

Visualizing the training and test-set result:
In the visualization process what we will do is, we will try to visualize the test and training set result .To do this we have to plot a graph for the decision tree classifier.

Training & Testing



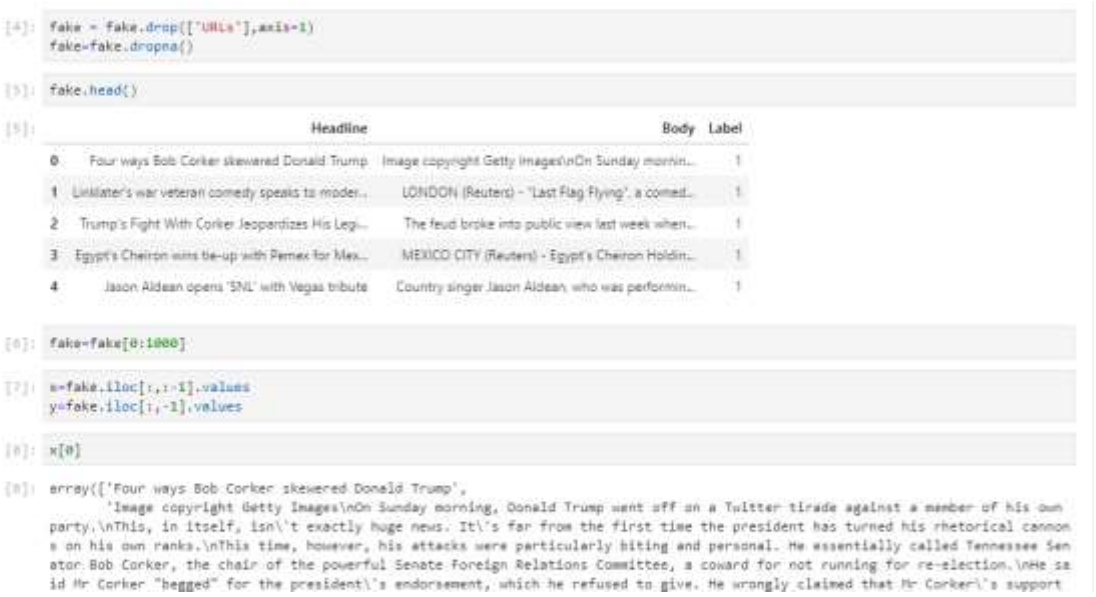
```
[1]: import pandas as pd
     import numpy as np
```

```
[2]: fake = pd.read_csv('data.csv')
```

```
[3]: fake.head()
```

[3]:

	URLs	Headline	Body	Label
0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1



[4]:
```
fake = fake.drop(['URLs'],axis=1)
fake=fake.dropna()
```

[5]:
```
fake.head()
```

[5]:

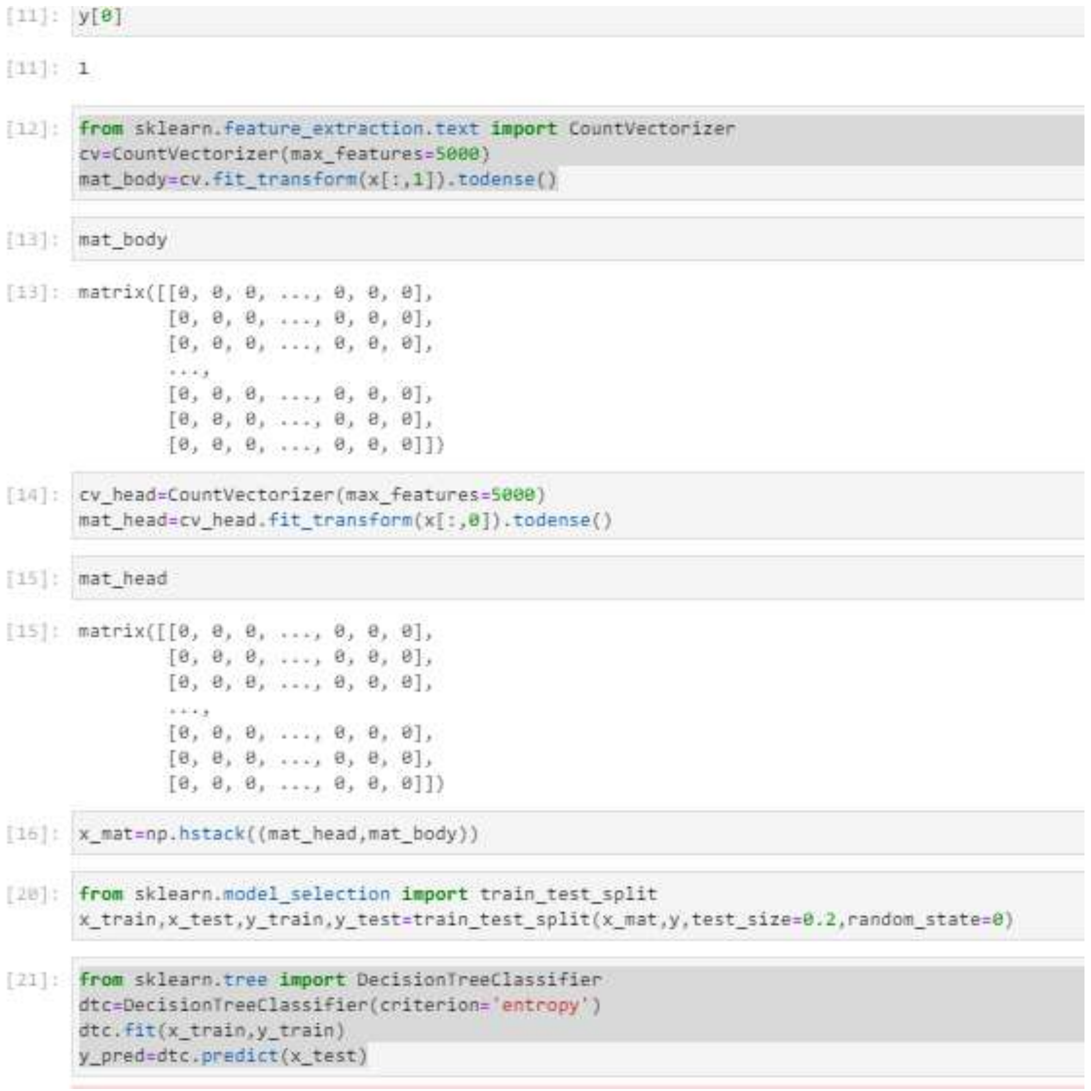
	Headline	Body	Label
0	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	Trump's Fight With Corker Jeopardizes His Leg...	The feud broke into public view last week when...	1
3	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1

[6]:
```
fake=fake[0:1000]
```

[7]:
```
x=fake.iloc[:,:-1].values
y=fake.iloc[:,-1].values
```

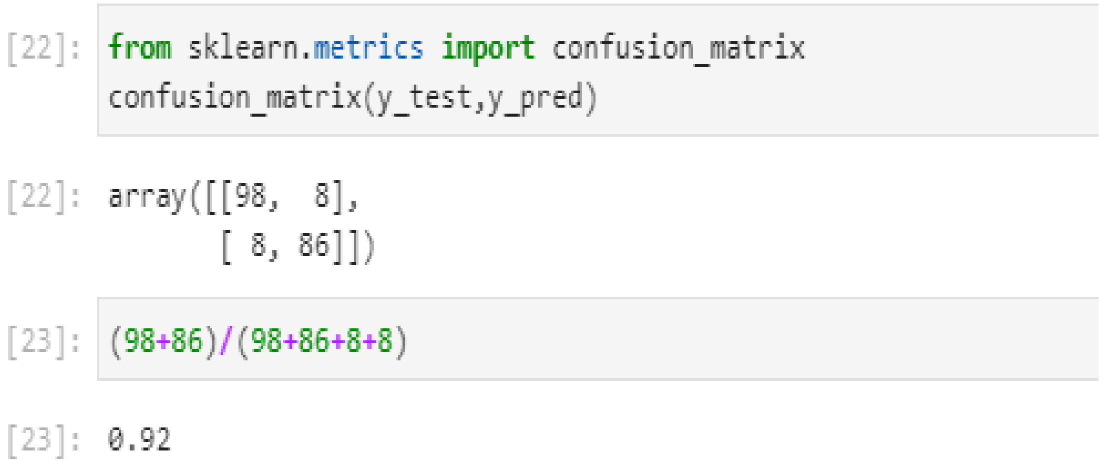
[8]:
```
x[0]
```

[8]:
```
array(['Four ways Bob Corker skewered Donald Trump',
       'Image copyright Getty Images\nOn Sunday morning, Donald Trump went off on a Twitter tirade against a member of his own
party.\nThis, in itself, isn\'t exactly huge news. It\'s far from the first time the president has turned his rhetorical cannon
s on his own ranks.\nThis time, however, his attacks were particularly biting and personal. He essentially called Tennessee Sen
ator Bob Corker, the chair of the powerful Senate Foreign Relations Committee, a coward for not running for re-election.\nHe sa
id Mr Corker "begged" for the president\'s endorsement, which he refused to give. He wrongly claimed that Mr Corker\'s support
```



```
[11]: y[0]

[11]: 1

[12]: from sklearn.feature_extraction.text import CountVectorizer
      cv=CountVectorizer(max_features=5000)
      mat_body=cv.fit_transform(x[:,1]).todense()

[13]: mat_body

[13]: matrix([[0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0],
              ...,
              [0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0]])

[14]: cv_head=CountVectorizer(max_features=5000)
      mat_head=cv_head.fit_transform(x[:,0]).todense()

[15]: mat_head

[15]: matrix([[0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0],
              ...,
              [0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0],
              [0, 0, 0, ..., 0, 0, 0]])

[16]: x_mat=np.hstack((mat_head,mat_body))

[20]: from sklearn.model_selection import train_test_split
      x_train,x_test,y_train,y_test=train_test_split(x_mat,y,test_size=0.2,random_state=0)

[21]: from sklearn.tree import DecisionTreeClassifier
      dtc=DecisionTreeClassifier(criterion='entropy')
      dtc.fit(x_train,y_train)
      y_pred=dtc.predict(x_test)
```

OUTPUT



```
[22]: from sklearn.metrics import confusion_matrix
      confusion_matrix(y_test,y_pred)

[22]: array([[98,  8],
             [ 8, 86]])

[23]: (98+86)/(98+86+8+8)

[23]: 0.92
```

CHAPTER 6

Conclusion:

The method of identifying the news manually needs a very good information of this domain and good experience to spot anomalies or errors in the given context. In this analysis, we are required to mention the matter of classifying or identifying the fake news articles with the help of machine learning model .The data we tend to employed in our work, we get it from online open source platform and it have the articles of news from a very large number of domains which covers the maximum amount of the news and also covers various domains such as political and sports news. The main idea of this research paper is identifying the pattern in the given information in text that is able to find the difference between the fake article news from the true news . detection of Fake news has several problems that needs attention of developers, data scientists, scholars and researchers. for example, in order the to reduce the fake news from spreading, important key components or steps concerned within the spread of article or story is a very important step in the way. Machine learning models and Graph theory are usually used to identify the key sources that are involved in the spread of fake articles and fake news.

REFERENCES:

IDEA:-https://olympus.greatlearning.in/courses/14365

Studied from:- https://www.javatpoint.in/machine-learning-decision-tree-classification-algorithm

https://www.researchgate.net/publication/339022255

S. B. Parikh, V. Patil, and P. K. Atrey, “On the Origin, Proliferation and Tone of Fake News,” Proc. - 2nd Int. Conf. Multimed. Inf. Process. Retrieval, MIPR 2019, pp. 135–140, 2019.