FAKE NEWS DETECTION

A Minor Project II Report
Submitted in Partial fulfillment for the award of
Bachelor of Technology in CSE-AIDS

Submitted to
RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA
BHOPAL (M.P)

MINOR PROJECT II REPORT
Submitted by
SMRITI VISHWAKARMA[0103AD223D04

Under the supervision of
DR. NIDHI SINGH
Assistant Prof.

Department of CSE-AIDS
Lakshmi Narain College of Technology, Bhopal (M.P.)
Session 2023-24

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY, BHOPAL

DEPARTMENT OF CSE-AIDS

CERTIFICATE

This is to certify that the work embodied in this project work entitled **”FAKE NEWS DETECTION”** has been satisfactorily completed by the **SMRITI VISHWAKARMA** (0103AD223D04.).It is a bonafide piece of work, carried out under the guidance in **Department of CSE-AIDS**, **Lakshmi Narain College of Technology, Bhopal** for the partial fulfillment of the **Bachelor of Technology** during the academic year 2023-24.

Approved By

{Dr. Nidhi Singh}
(Assistant Prof.)
(Guide)

Dr. Tripti Saxena
Prof. & Head
Department of CSE-AIDS

LAKSHMI NARAIN COLLEGE OF TECHNOLOGY, BHOPAL

DEPARTMENT OF CSE-AIDS

ACKNOWLEDGEMENT

We express our deep sense of gratitude to Prof Dr. Nidhi Singh. (Guide) department of CSE- AIDS L.N.C.T., Bhopal. Whose kindness valuable guidance and timely help encouraged me to complete this project.

A special thank goes to Dr. Tripti Saxena (HOD) who helped me in completing this project work. He exchanged his interesting ideas & thoughts which made this project work successful.

We would also thank our institution and all the faculty members without whom this project work would have been a distant reality.

SMRITI VISHWAKARMA[0103AD223D04]

INDEX

Total Approx. 50-60 Pages

CHAPTER 1

PROBLEM DOMAIN

I. Introduction

The effects of fake news have increased exponentially in the recent past and something must be done to prevent this from continuing in the future. The dangerous effects of fake news, as previously defined, are made clear by events such as in which a man attacked a pizzeria due to a widespread fake news article. This story along with analysis provides evidence that humans are not very good at detecting fake news, possibly not better than chance. As such, the question remains whether machines can do a better job. A machine can solve the fake news problem using supervised learning that extracts feature of the language and content only within the source in question, without utilizing any fact-checker or knowledge base. Do you trust all the news you hear from social media? All news is not real, right? So how will you detect the fake news? – The answer is Python. By practicing such an advanced python project of detecting fake news, we will easily make a difference between real and fake news. Before moving ahead in this advanced Python project, we have to be aware of related terms of fake news like the TF-IDF Vectorizer, and the Passive Aggressive Classifier. This project will work you through the necessary steps and techniques used to implement such an analysis.

1.1.Problem Statement

This project proposes the question of whether it is possible to detect fake news through machine learning models. Specifically, the aim of this project is to determine the ideal model that is efficient in predicting fake news while also limiting the cost of memory and storage for computation. ”Fake news” has been a very recent and prevalent problem within recent years.

I.2 Detail of the Problem Application

1.2.1 Area & Domain

Fake news spreads like a wildfire and this is a big issue in this era. We can learn how to distinguish fake news from real one. We will be using a supervised learning approach to implement the model.

As consequence of the increase in cases of fake news in recent years, efforts have been made to crack down on the spread of misinformation throughout social media platforms. All popular social media platforms (Facebook, Twitter, Spotify, and YouTube) have permanently banned Alex Jones from using their networks following the events of ”Pizza Gate” in addition to multiple questionable accusations made by Jones, including an accusation made by Jones claiming that the Sandy Hook shooting was ”faked”. Despite efforts of many social media websites and governments cracking down on fake news, many young people today generally are not able to tell the difference between fake news and real news. According to a Stanford study, it found that many students have a very strong inability in discerning between fake news. In the study, high school students were given two posts announcing the candidacy of Donald Trump’s presidential campaign. One post was given by an actual Fox News account another one posted by an account that ”looked” like it was from Fox News. 25% of the could not tell the difference between real and fake news sources. With over 30% of students favoring that the fake news account was more trustworthy. Indeed, some politically charged or bogus articles that would be esteemed false frequently have more perspectives and offers via web-based media destinations than real news stories towards the most recent three months of the political race. As per an examination by Buzz-channel, posts, and stories composed from the best twenty most noteworthy performing trick locales and hyper-hardliners had over 8.7 million offers, ”responses” and remarks contrasted with the main twenty most noteworthy performing significant news associations had about 7.4 million offers, ”responses” and remarks via online media destinations.

The research problem was initially defined through the following use cases: In light of a single event/story, the framework would decide whether certain sources or articles are regarded to be fake news dependent on a given likelihood.

Through the sources analyzed the machine learning agent would assign a level of bias and factuality of these articles by comparing them to each other and assign scores of the bias and factuality of the source.

A sort of sensationalist reporting, counterfeit news exemplifies bits of news that might be scams and is commonly spread through web-based media and other online media. This is regularly done to further or force certain thoughts and is frequently accomplished with political plans. Such news things may contain bogus and additionally overstated cases and may wind up being viral by calculations, and clients may wind up in a channel bubble.

I.3 Challenges and Motivation

1.3.1 Motivation—

The motivation for research on this topic was that this is a relatively new area of research with many
opinions but not many concrete solutions. Many implementations focus primarily on the host of the article, but even articles hosted on otherwise trustworthy websites can be classified as fake news. The primary motivation of this project was to bring awareness, propose a solution, and work towards minimizing the effects of fake news.

1.3.2 Challenges—

Throughout the project and its analysis study, we faced some challenges such as:

- **Data Collection**: While collecting the data, we faced an issue with the relation of topics among news feeds. This is due to the variety of categorized topics covered by most informative platforms.
- **Data Analysis and Interpretation**: After collecting these data, we spend a lot of time analyzing from a feature-based perspective.
- **Data Preprocessing Methodology**: After proper analysis and interpretation, we need to prepare and sanitize the data for it to be suitable for more insights into the learning process.
- **Learning Model Selection** (Which model suits well the given problem.)
- **Update the Model for better Accuracy**: Here we try to tune the parameter

CHAPTER 2

LITERATURE SURVEY

The available literature has described many automatic detection techniques of fake news and deception posts. Since there are multidimensional aspects of fake news detection ranging from using chatbots for spread of misinformation to use of clickbaits for the rumor spreading There are many clickbaits available in social media networks including facebook which enhance sharing and liking Proceedings of posts which in turn spreads falsified information. Lot of work has been done to detect falsified information.

2. MEDIA RICH FAKE NEWS DETECTION: A SURVEY

In general, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4% accuracy using logistic classifier**.**

2.1 WEAKLY SUPERVISED DETECTION ON TWITTERLEARNING FOR FAKE NEWS

The problem of automatic detection of fake news in social media, e.g., on Twitter,has recently drawn some attention. Although, from a technical perspective, it can be regarded as a straight-forward, binary classification problem, the major challenge is the collection of large enough training corpora, since manual annotation of tweets as fake or non-fake news is an expensive and tedious endeavor. In this paper, we discuss a weakly supervised approach, which automatically collects a large-scale,but very noisy training dataset comprising hundreds of thousands of tweets. During collection, we automatically label tweets by their source, i.e., trustworthy or untrustworthy source, and train a classifier on this dataset. We then use that classifier for a different classification target, i.e., the classification of fake and non-fake tweets. Although the labels are not accurate according to the new classification target (not all tweets by an untrustworthy source need to be fake news, and vice versa), we show that despite this unclean inaccurate dataset, it is possible to detect fake news with an F1 score of up to 0.9

2.2 FAKE NEWS DETECTION IN SOCIAL MEDIA

Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers". Social media and news outlets publish fake news to increase readership or as part of psychological warfare. Ingeneral, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4% accuracy using logistic classifier. Automatic Online Fake News Detection Combining Content and Social Signals The proliferation and rapid diffusion of fake news on the Internet highlight the need of automatic hoax detection systems. In the context of social networks, machine learning (ML) methods can be used for this purpose. Fake news detection strategies are traditionally either based on content analysis (i.e. analyzing the content of the news) or - more recently - on social context models, such as mapping the news‟diffusion pattern. In this paper, we first propose a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing their already high accuracy by up

to 4.8%. Second, we implement our method within a Facebook Messenger chatbot and validate it with a real-world application, obtaining a fake news detection accuracy of 81.7%.
In recent years, the reliability of information on the Internet has emerged as a crucial issue of modern society. Social network sites (SNSs) have revolutionized the way in which information is spread by allowing users to freely share content. As a consequence, SNSs are also increasingly used as vectors for the diffusion of misinformation and hoaxes. The amount of disseminated information and the rapidity of its diffusion make it practically impossible to assess reliability in a timely manner, highlighting the need for automatic hoax detection systems. As a contribution towards this objective, we show that Facebook posts can be classified with high accuracy as hoaxes or non-hoaxes on the basis of the users who "liked" them. We present two classification techniques, one based on logistic regression, the other on a novel adaptation of Boolean crowdsourcing algorithms. On a dataset consisting of 15,500 Facebook posts and 909,236 users, we obtain classification accuracies exceeding 99% even when the training set contains less than 1% of the posts. We further show that our techniques are robust: they work even when we restrict our attention to the users who like both hoax and non-hoax posts. These results suggest that mapping the diffusion pattern of information can be a useful component of automatic hoax detection systems.

2.3 MISLEADING ONLINE CONTENT

Tabloid journalism is often criticized for its propensity for exaggeration,sensationalization, scare-mongering, and otherwise producing misleading and low-quality news. As the news has moved online, a new form of tabloidization has emerged: „click baiting.‟ „Clickbait‟ refers to “content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page”[„clickbait,‟ n.d.] and has been implicated in the rapid spread of rumor and misinformation online. This paper examines potential methods for the automatic detection of clickbait as a form of deception. Methods for recognizing both textual and non-textual click baiting cues are surveyed, leading to the suggestion that a hybrid approach may yield best results.
Big Data Analytics and Deep Learning are two high-focus of data science.Big Data has become important as many organizations both public and private have been collecting

massive amounts of domain-specific information, which can contain useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics. Companies such as Google and Microsoft are analyzing large volumes of data for business analysis and decisions, impacting existing and future technology. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. Complex abstractions are learnt at a given level based on relatively simpler abstractions formulated in the preceding level in the hierarchy.

A key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabeled and un-categorized. In the present study, we explore how Deep Learning can be utilized for addressing some important problems in Big Data Analytics, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. We also investigate some aspects of Deep Learning research that need further exploration to incorporate specific challenges introduced by Big Data Analytics, including streaming data, high-dimensional data, scalability of models, and distributed computing. We conclude by presenting insights into relevant future works by posing some questions,including defining data sampling criteria, domain adaptation modeling, defining criteria for obtaining useful data abstractions, improving semantic indexing, semi -supervised learning, and active learning.

CHAPTER 3

MAJOR OBJECTIVE & SCOPE OF PROJECT

3. OBJECTIVE

The objective of this project is to examine the problems and possible significances related with the spread of fake news. We will be working on different fake news data set in which we will apply different machine learning algorithms to train the data and test it to find which news is the real news or which one is the fake news. As the fake news is a problem that is heavily affecting society and our perception of not only the media but also facts and opinions themselves. By using the artificial intelligence and the machine learning, the problem can be solved as we will be able to mine the patterns from the data to maximize well defined objectives. So, our focus is to find which machine learning algorithm is best suitable for what kind of text dataset. Also, which dataset is better for finding the accuracies as the accuracies directly depends on the type of data and the amount of data. The more the data, more are your chances of getting correct accuracy as you can test and train more data to find out your results.

3.1 SCOPE OF PROJET

The scope of fake news detection is broad and continuously evolving due to the dynamic nature of misinformation and disinformation. Here are some key aspects that fall within the scope of fake news detection:

1. **Content Analysis**: Analyzing the content of news articles, social media posts, images, and videos to identify misleading or fabricated information. This involves examining language, tone, sources, and context to assess the credibility of the information.

2. **Source Verification**: Verifying the credibility of the sources cited in news stories or social media posts. This includes assessing the reputation and reliability of news outlets, journalists, and individuals sharing information.
3. **Fact-Checking**: Conducting fact-checking procedures to verify the accuracy of claims made in news articles or social media posts. Fact-checkers rely on evidence-based research and reputable sources to debunk false or misleading information.
4. **Contextual Analysis**: Considering the broader context in which information is shared to understand its potential impact and implications. This involves examining social, political, and cultural factors that may influence the spread of fake news.
5. **Technological Solutions**: Developing and implementing automated tools and algorithms to detect and flag fake news content. These solutions may use natural language processing, machine learning, and data mining techniques to analyze large volumes of information and identify patterns indicative of misinformation.
6. **User Education**: Educating individuals about the dangers of fake news and providing them with critical thinking skills to evaluate information critically. This includes raising awareness about common misinformation tactics and encouraging skepticism towards sensationalist or unverified claims.
7. **Collaborative Efforts**: Collaborating with journalists, fact-checking organizations, tech companies, and academic institutions to share resources, best practices, and methodologies for combating fake news. Collective action is essential for addressing the spread of misinformation across different platforms and channels.
8. **Legal and Policy Interventions**: Implementing legal and policy measures to hold purveyors of fake news accountable and mitigate its harmful effects. This may include regulations governing the dissemination of false information and penalties for those who intentionally deceive the public.
9. **Ethical Considerations**: Addressing ethical concerns related to fake news detection, such as privacy implications, bias in algorithmic decision-making, and the potential for censorship. It's essential to strike a balance between protecting against misinformation and upholding principles of free speech and information access.

CHAPTER 4

PROBLEM ANALYSIS AND REQUIREMENT SPECIFICATION

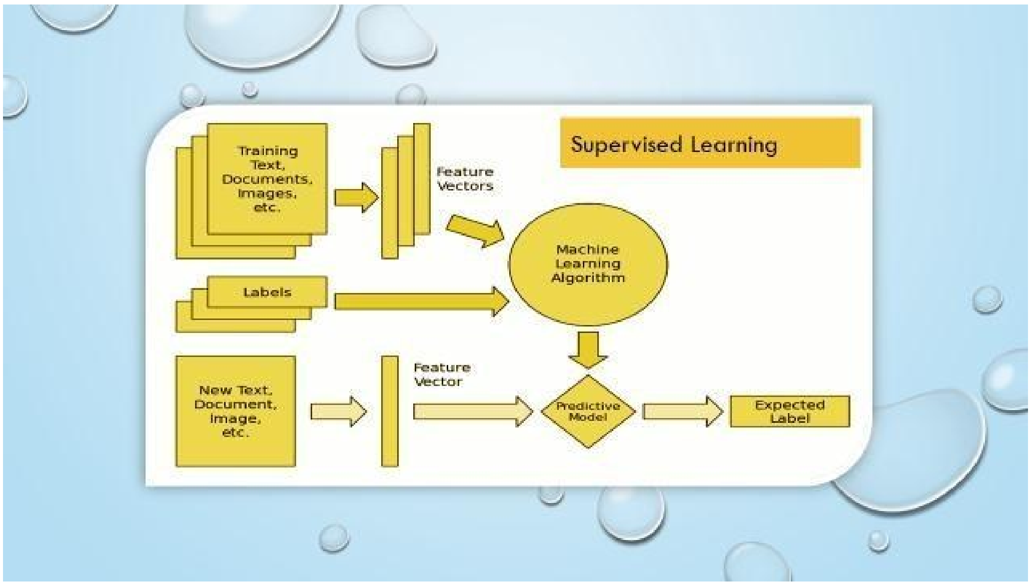
4. EXISTING SYSTEM

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature. Conroy, Rubin, and Chen outlines several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora

4.1 PROPOSED SYSTEM

In this paper a model is build based on the count vectorizer or a tfidf matrix (i.e.) word tallies relatives to how often they are used in other artices in your dataset) can help . Since this problem is a kind of text classification, Implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now the9 next step is to extract the most optimal features for count vectorizer or tfidf-vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset

4.2 SYSTEM ARCHITECTURE


Supervised Learning
Training Text, Documents, Images, etc.
Feature Vectors
Labels
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

4.3 REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analysing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

4.4 FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

- **Usability**

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

- **Robustness**24

It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

- **Security**

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

- **Reliability**

It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

- **Compatibility**

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

- **Flexibility**

The flexibility of the project is provided in such a way that is has the ability to run on different environments being executed by different users.

- **Safety**

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one‟s personal information.

- **Compatibility**

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience

.

- **Flexibility**

The flexibility of the project is provided in such a way that is has the ability to run on different environments being executed by different users.

- **Safety**

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one‟s personal information.

4.6 NON- FUNCTIONAL REQUIREMENTS

- **Portability**

It is the usability of the same software in different environments. The project can be run in any operating system.

- **Performance**

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

- **Accuracy**

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

- **Maintainability**

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

4.6 SYSTEM DESIGN AND TESTING PLAN

4.6.1INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:26

[12] What data should be given as input?

[13] How the data should be arranged or coded?

[14] The dialog to guide the operating personnel in providing input.

[15]Methods for preparing input validations and steps to follow when error occur.

4.6.2OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs.

In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system‟s relationship to help user decision-making. The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

4.7 SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4.8 TEST PROCEDURE

4.8.1SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product

It is the process of exercising software with the intent of ensuring that the
Software system meets its requirements and user expectations and does not fail in
an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement

4.8.2UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

4.8.3 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

4.8.4FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

- **Valid Input**: identified classes of valid input must be accepted.
- **Invalid Input**: identified classes of invalid input must be rejected.

- **Function**: identified functions must be exercised.
- **Output**: identified classes of application outputs must be exercised.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

4.8.5WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level. Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works

4.8.6 ACCEPTANCE TESTING

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

4.9 FLASK FRAMEWORK

Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Poccoprojects. Http protocol is the foundation of data communication in world wide web. Different methods of data retrieval from specified URL are defined in this protocol.

The following table summarizes different http methods −

S.r.No.	Methods	Description
1	**GET**	Sends data in unencrypted form to the server. Most common method.
2	**HEAD**	Same as GET, but without response body
3	**POST**	Used to send HTML form data to server. Data received by POST method is not cached by server.
4	**PUT**	Replaces all current representations of the target resource with the uploaded content
5	**DELETE**	Removes all current representations of the target resource given by a UR

By default, the Flask route responds to the **GET** requests. However, this preference can be altered by providing methods argument to **route()** decorator.
In order to demonstrate the use of **POST** method in URL routing, first let us create an HTML form and use the **POST** method to send form data to a URL

4.9.1LOGIN PAGE

Form data is POSTed to the URL in action clause of form tag.

http://localhost/loginis mapped to the **login()** function. (FIG4.1)Since the server has received data by **POST** method, value of „nm" parameter obtained from the form data is obtained by

user = request. form['nm']

4.9.2DASHBOARD

Change the method parameter to **„GET"** in **login.html** and open it again in the browser. The data received on server is by the **GET** method.(FIG 4.2) The value of „nm" parameter is now obtained by

User = request.args.get(„nm‟)

Here, **args** is dictionary object containing a list of pairs of form parameter and its corresponding value. The value corresponding to „nm‟ parameter is passed on to „/success‟ URL as before.

CHAPTER 5

DETAILED DESIGN
(MODELING AND ERD/DFD)

5. DATA FLOW DIAGRAM

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

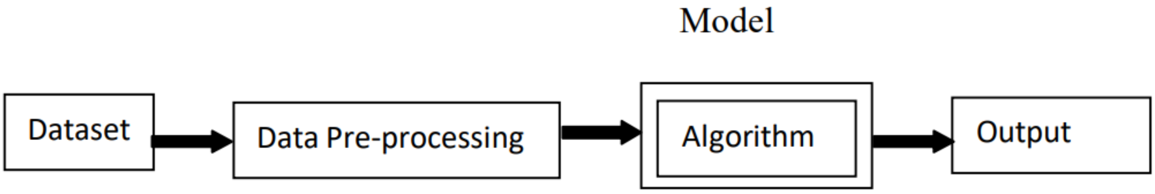
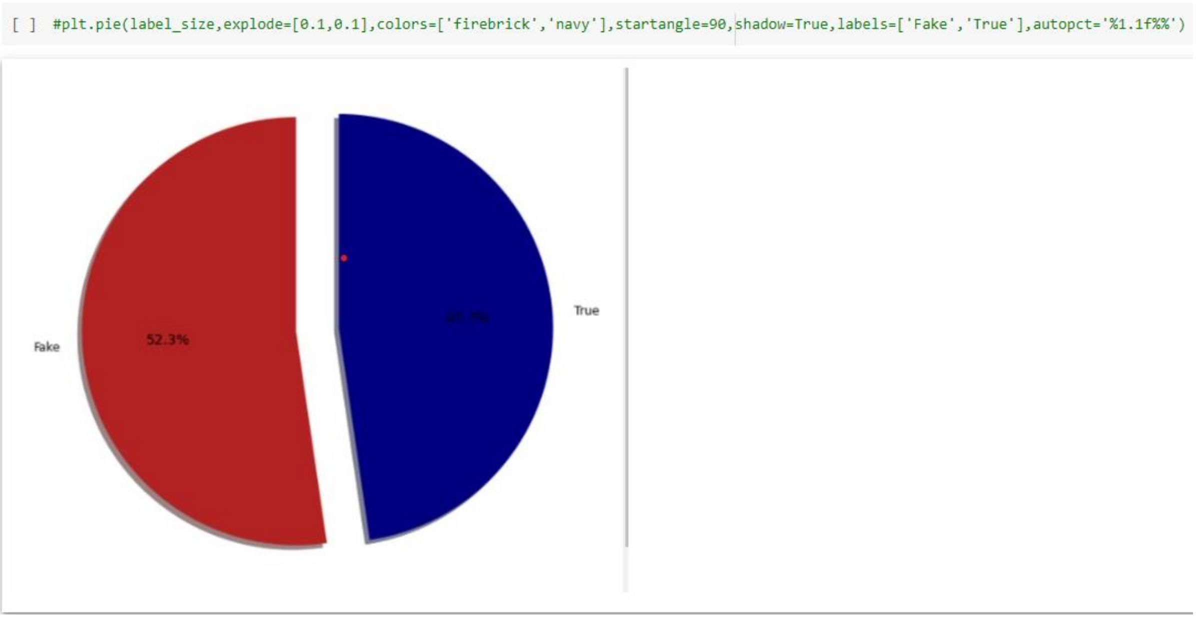


Fig.5



```
[ ]  #plt.pie(label_size,explode=[0.1,0.1],colors=['firebrick','navy'],startangle=90,shadow=True,labels=['Fake','True'],autopct='%1.1f%%')
```

Fig. 5.1: Comparing Fake and True Dataset

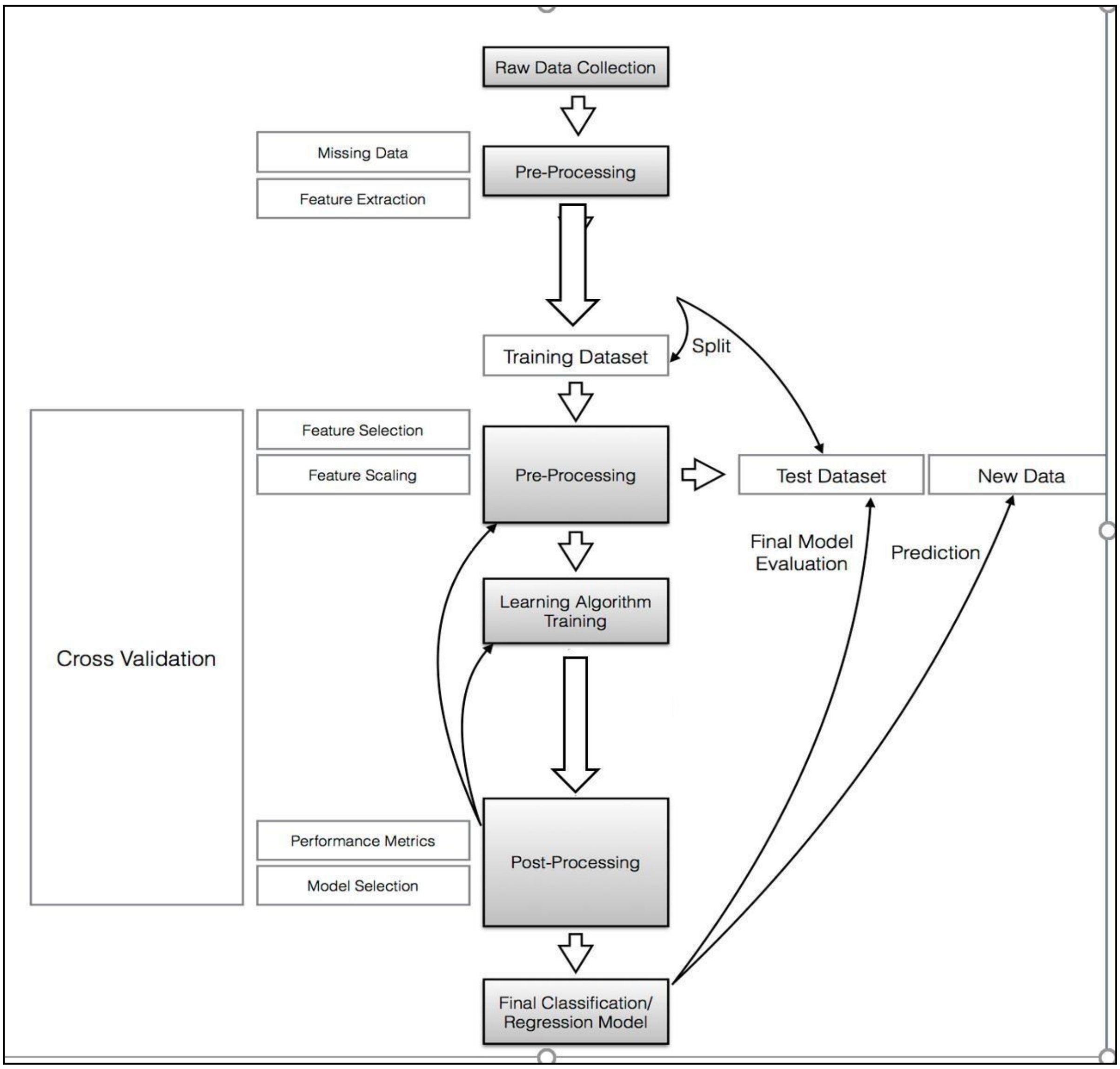
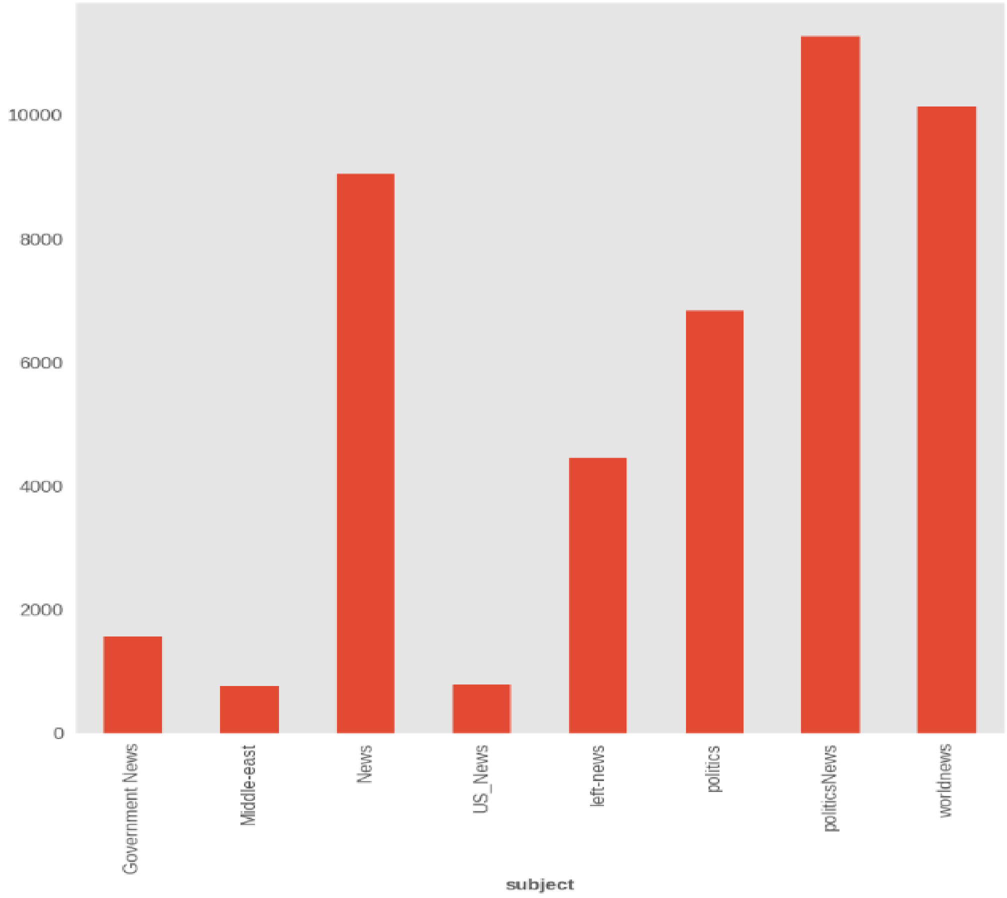
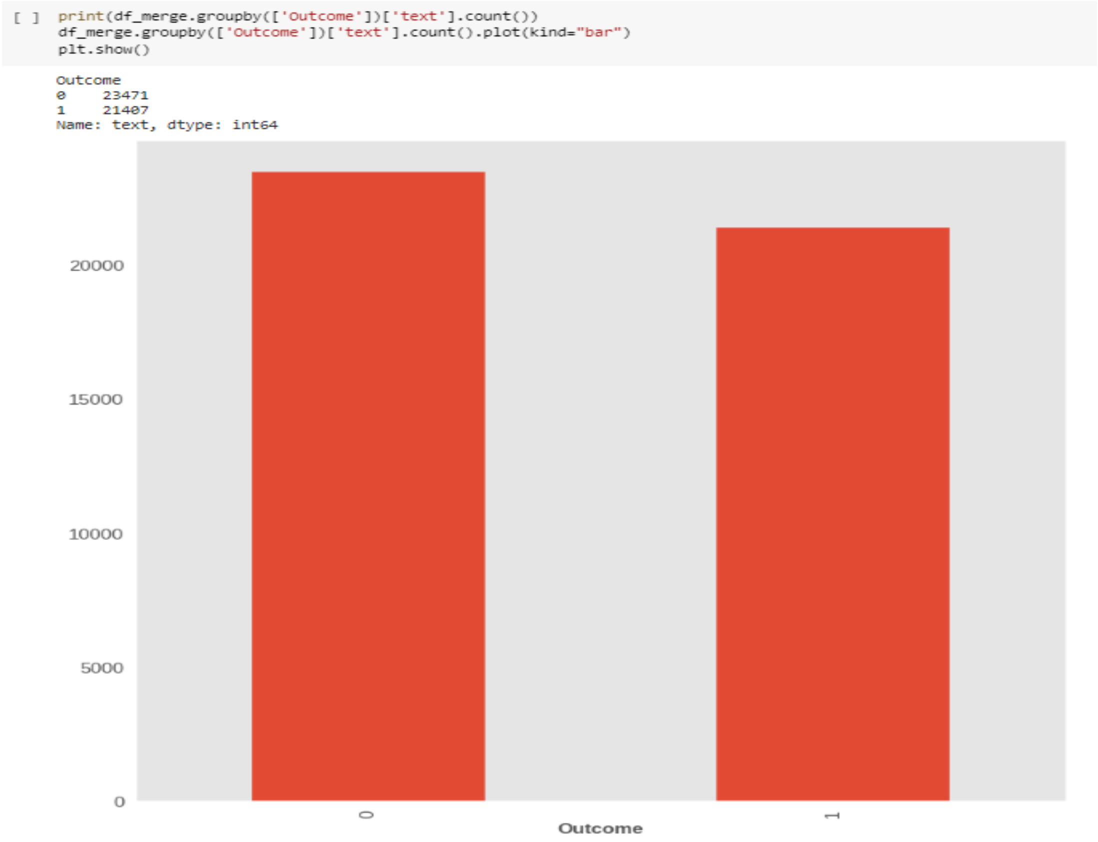

Raw Data Collection
Missing Data
Feature Extraction
Pre-Processing
Split
Training Dataset
Cross Validation
Feature Selection
Feature Scaling
Pre-Processing
Test Dataset
New Data
Final Model Evaluation
Prediction
Learning Algorithm Training
Performance Metrics
Model Selection
Post-Processing
Final Classification/ Regression Model

Fig 5.2 DFD



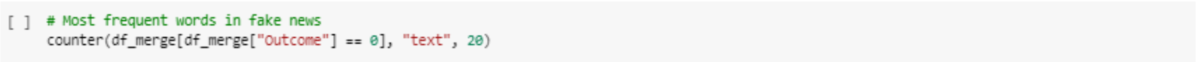
Graph 5.3: Frequency of subject of the news



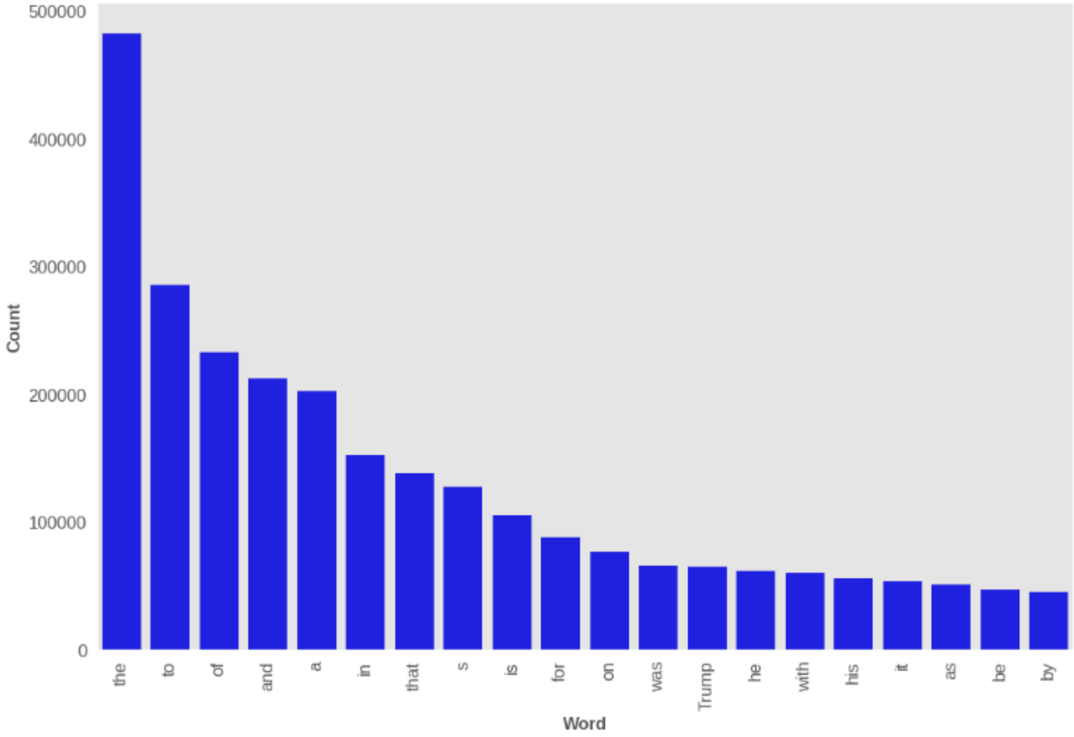
```
[ ]  print(df_merge.groupby(['Outcome'])['text'].count())
     df_merge.groupby(['Outcome'])['text'].count().plot(kind="bar")
     plt.show()
```

```
Outcome
0    23471
1    21407
Name: text, dtype: int64
```

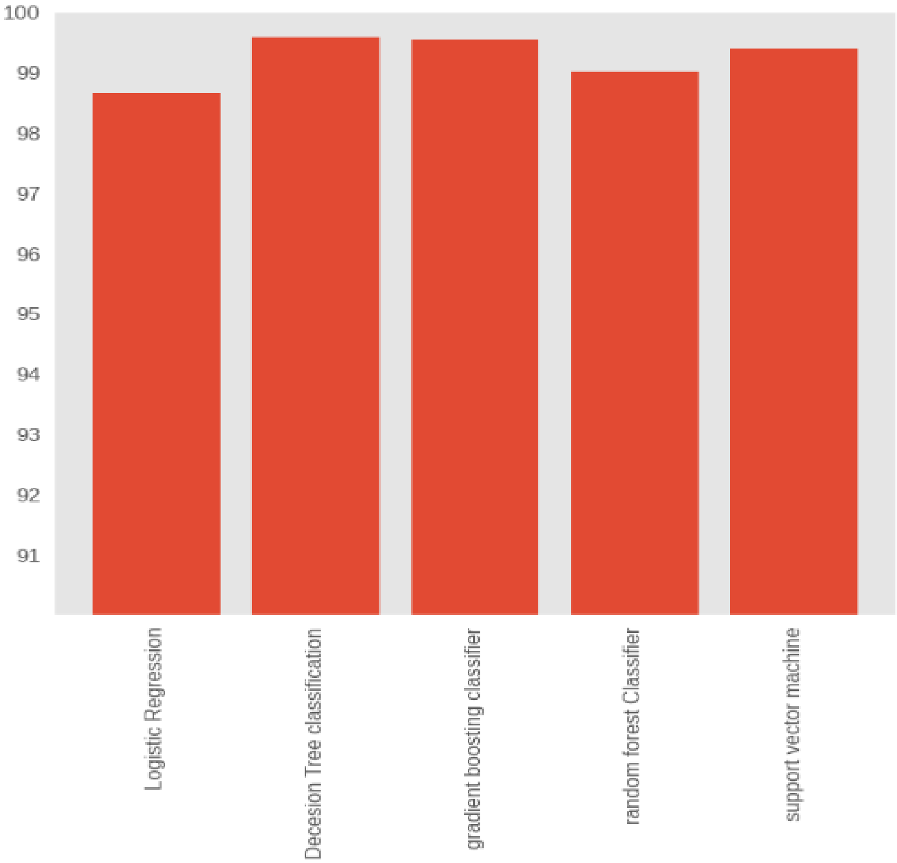
Graph 5.4.: Fake and Real News



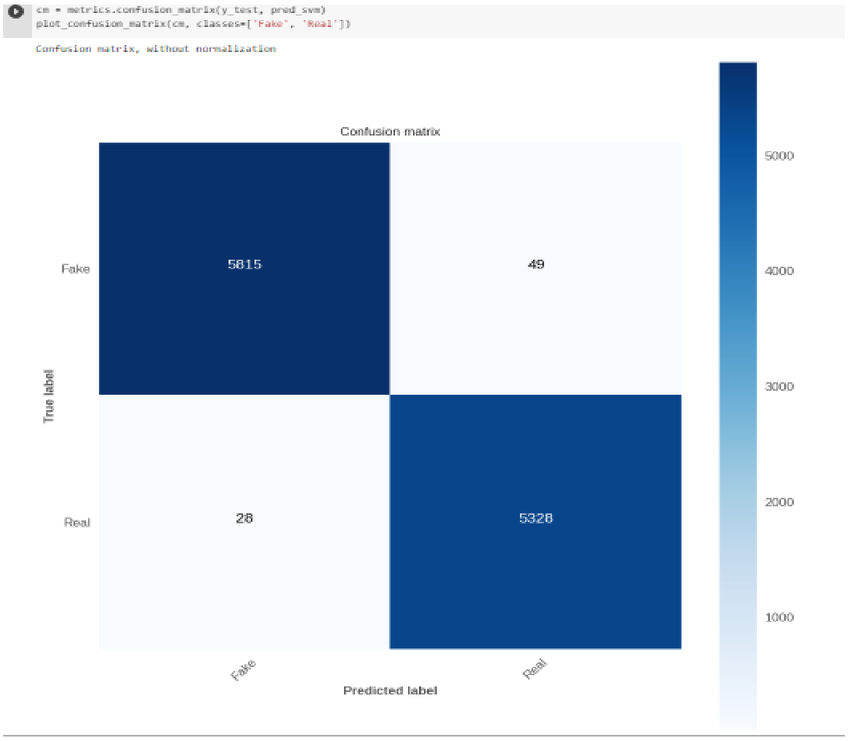
```
# Most frequent words in fake news
counter(df_merge[df_merge["Outcome"] == 0], "text", 20)
```



Graph 5.5: Frequency of words in fake news



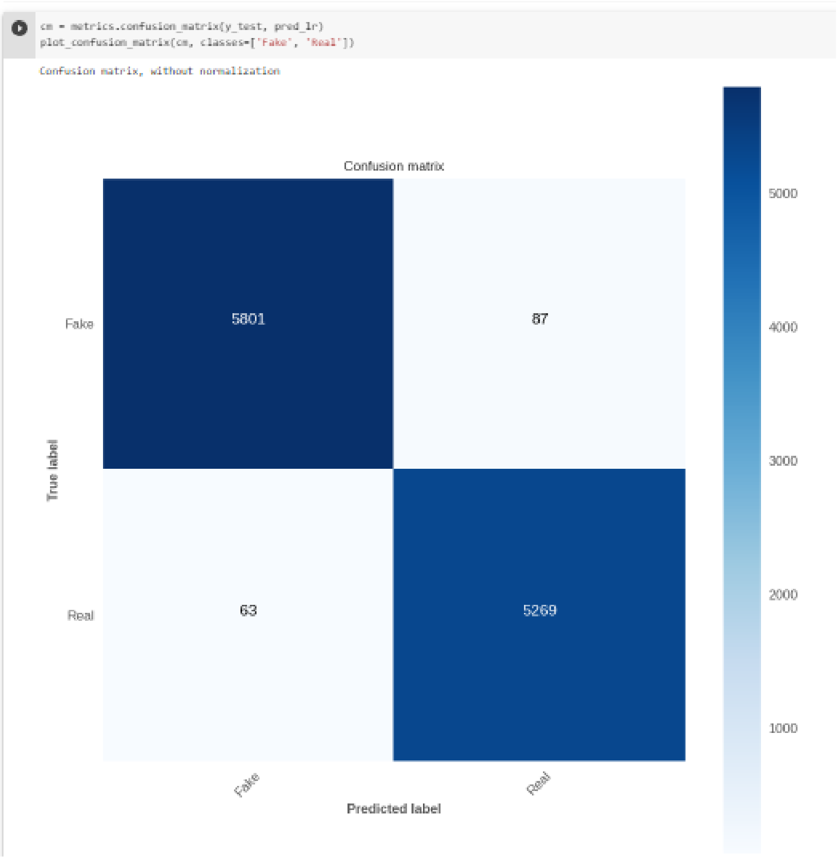
Graph : 5.6 Comparison of the accuracies of different models



```
cm = metrics.confusion_matrix(y_test, pred_svm)
plot_confusion_matrix(cm, classes=['Fake', 'Real'])
```

```
Confusion matrix, without normalization
```

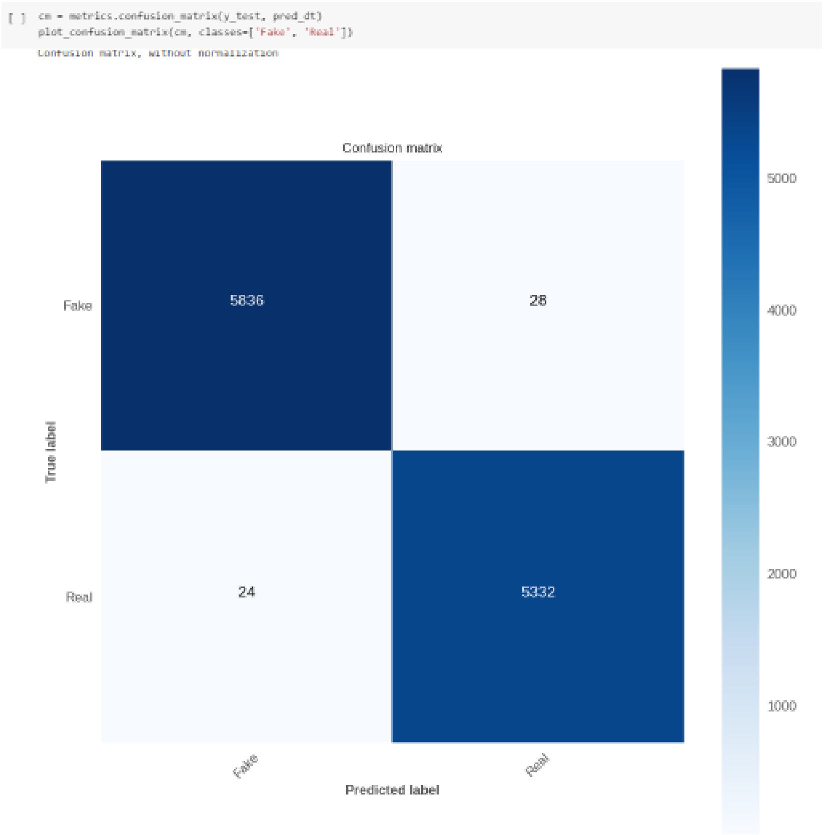
Fig. 8.3.2: Confusion Matrix from Support Vector Machine



```
cm = metrics.confusion_matrix(y_test, pred_lr)
plot_confusion_matrix(cm, classes=['Fake', 'Real'])
```

```
Confusion matrix, without normalization
```

Fig. 8.3.3: Confusion matrix from Logistic Regression

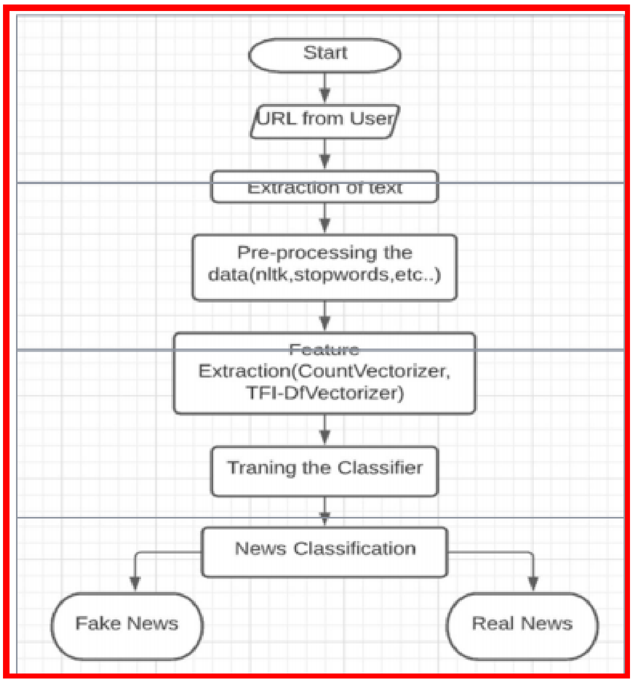


```
cm = metrics.confusion_matrix(y_test, pred_dt)
plot_confusion_matrix(cm, classes=['Fake', 'Real'])
```

```
Confusion matrix, without normalization
```

Fig .8.3.4: Confusion Matrix from Decision Tree Classification

FLOW CHART



Start
URL from User
Extraction of text
Pre-processing the data(nltk,stopwords,etc..)
Feature Extraction(CountVectorizer, TFI-DfVectorizer)
Traning the Classifier
News Classification
Fake News
Real News

CHAPTER 6

Hardware/Software Platform Environment

SYSTEM REQUIREMENTS

HARDWARE REQUIREMENTS:

- System - Pentium-IV
- Speed - 2.4GHZ
- Hard disk - 40GB
- Monitor - 15VGA color
- RAM - 512MB

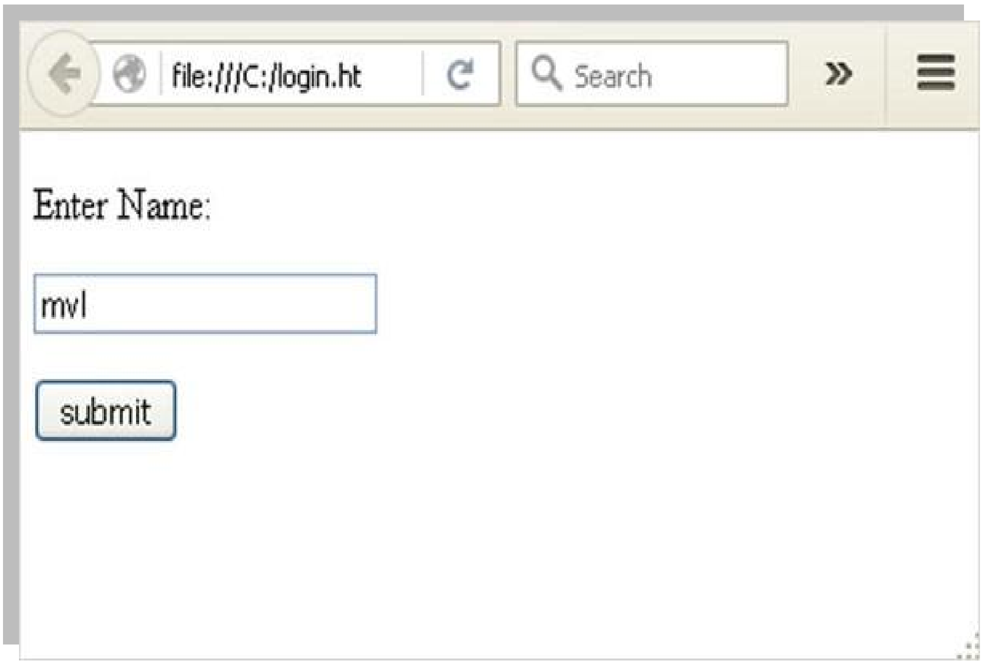
SOFTWARE REQUIREMENTS:

- Operating System - Windows XP
- Coding language – PYTHON

CHAPTER 7

SNAPSHOT OF INPUT & OUTPUT

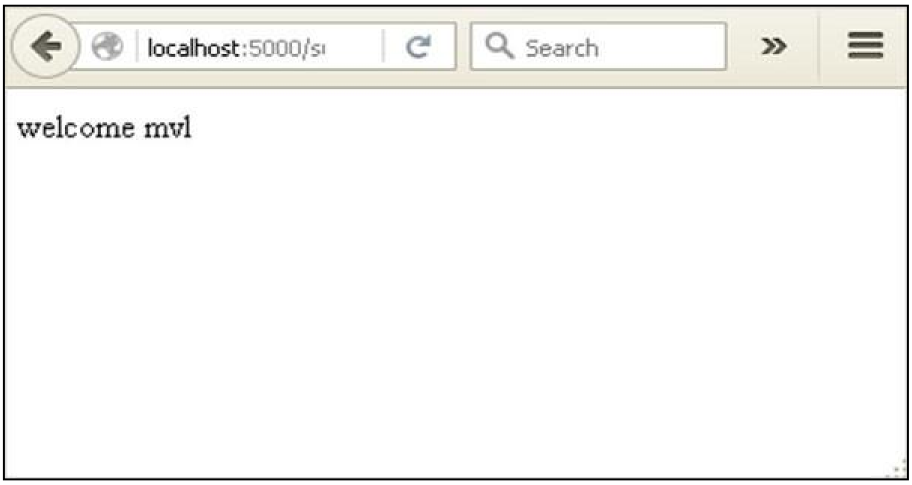
7. SCREENSHOT



Enter Name:

mvl

Fig-4.9.1 LOGIN PAGE



welcome mvl

Fig-4.9.DASHBOARD

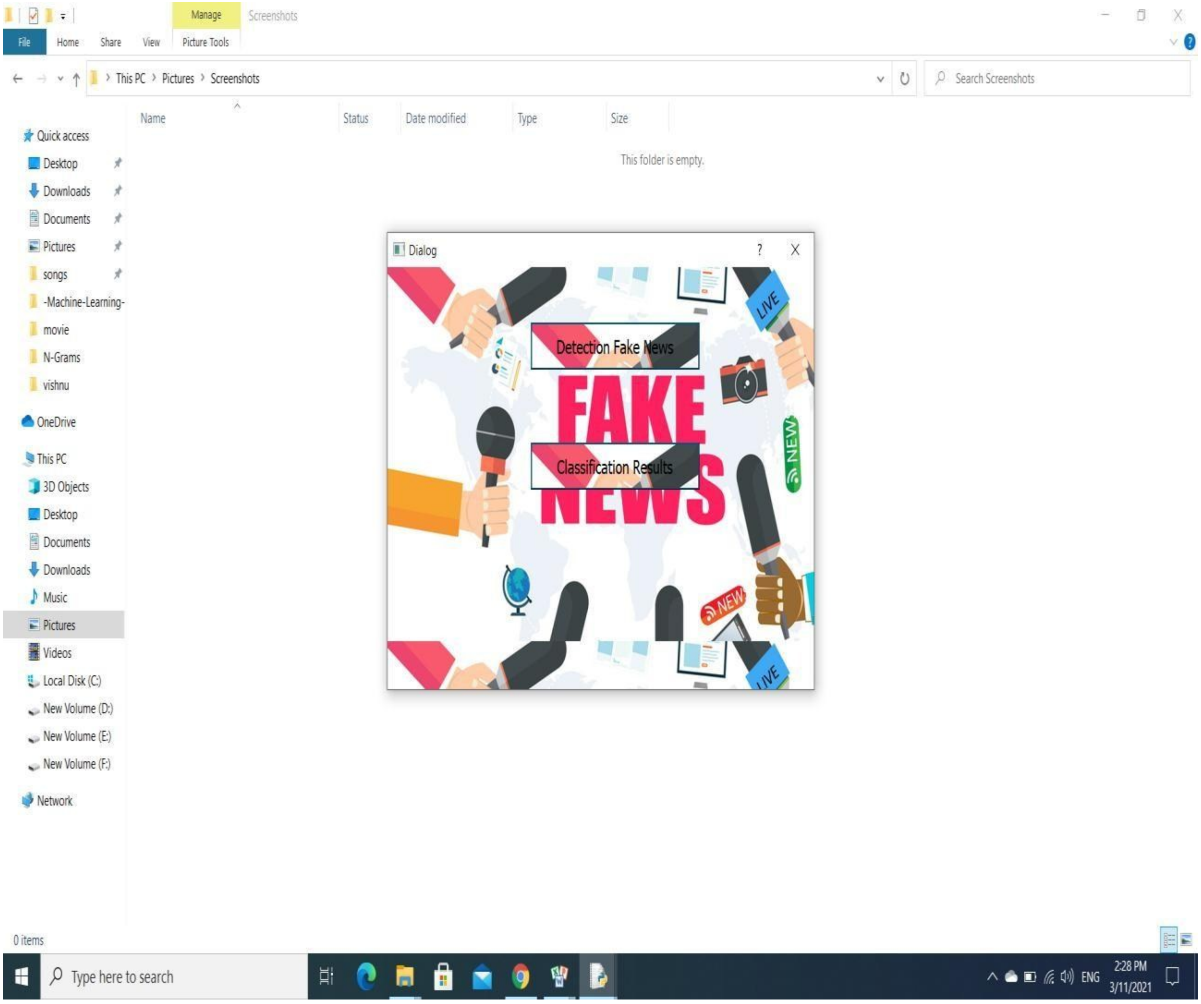
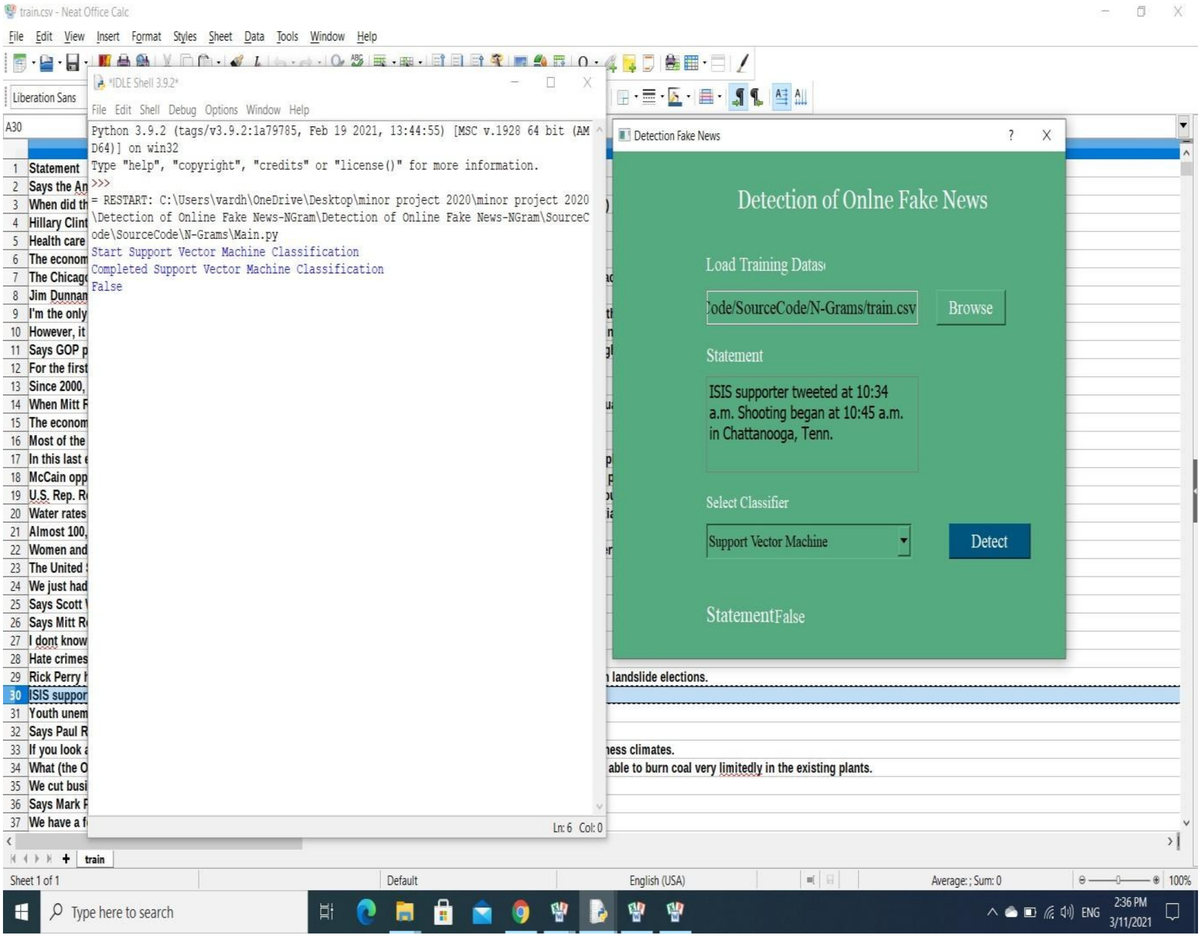

Manage
Screenshots
File
Home
Share
View
Picture Tools
This PC › Pictures › Screenshots
Search Screenshots
Name
Status
Date modified
Type
Size
This folder is empty.
Quick access
Desktop
Downloads
Documents
Pictures
songs
-Machine-Learning-
movie
N-Grams
vishnu
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
New Volume (D:)
New Volume (E:)
New Volume (F:)
Network
Dialog
Detection Fake News
FAKE
Classification Results
NEWS
LIVE
NEW
0 items
Type here to search
ENG
2:28 PM
3/11/2021

Fig.8.1



```
Python 3.9.2 (tags/v3.9.2:1a79785, Feb 19 2021, 13:44:55) [MSC v.1928 64 bit (AM
D64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:\Users\vardh\OneDrive\Desktop\minor project 2020\minor project 2020
\Detection of Online Fake News-NGram\Detection of Online Fake News-NGram\SourceC
ode\SourceCode\N-Grams\Main.py
Start Support Vector Machine Classification
Completed Support Vector Machine Classification
False
```

Detection of Onlne Fake News

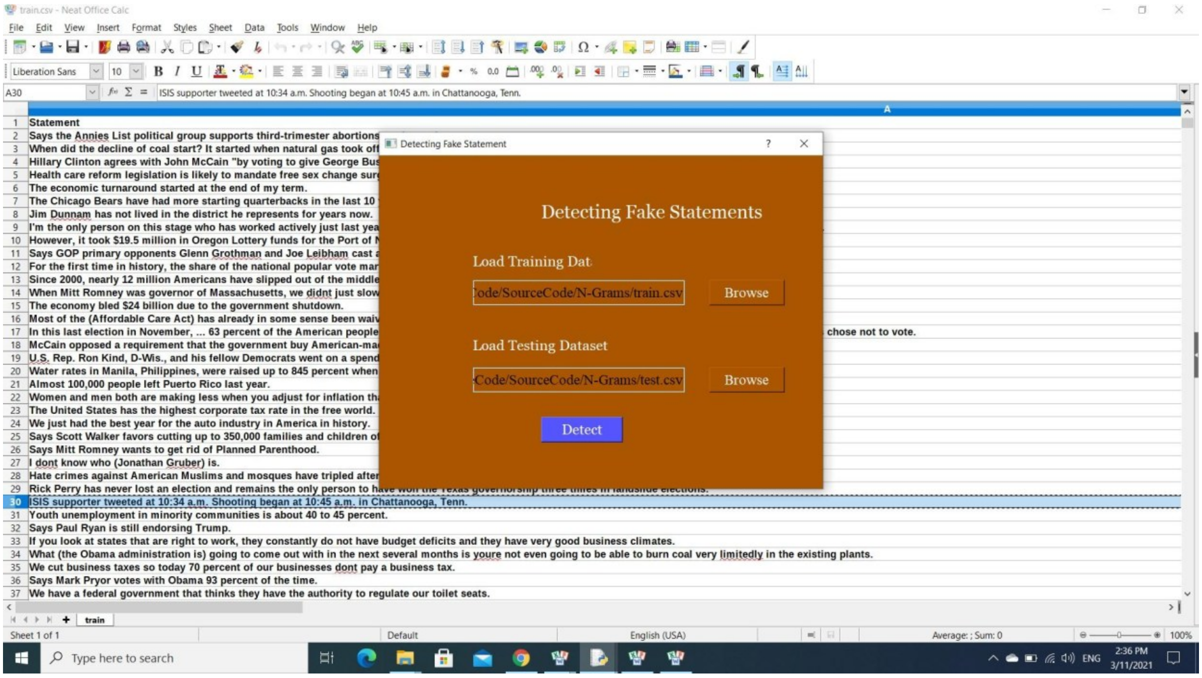
Load Training Datas: ‹ode/SourceCode/N-Grams/train.csv

Statement: ISIS supporter tweeted at 10:34 a.m. Shooting began at 10:45 a.m. in Chattanooga, Tenn.

Select Classifier: Support Vector Machine

StatementFalse

Fig:8.1 Checking statement with Dataset



A30: ISIS supporter tweeted at 10:34 a.m. Shooting began at 10:45 a.m. in Chattanooga, Tenn.

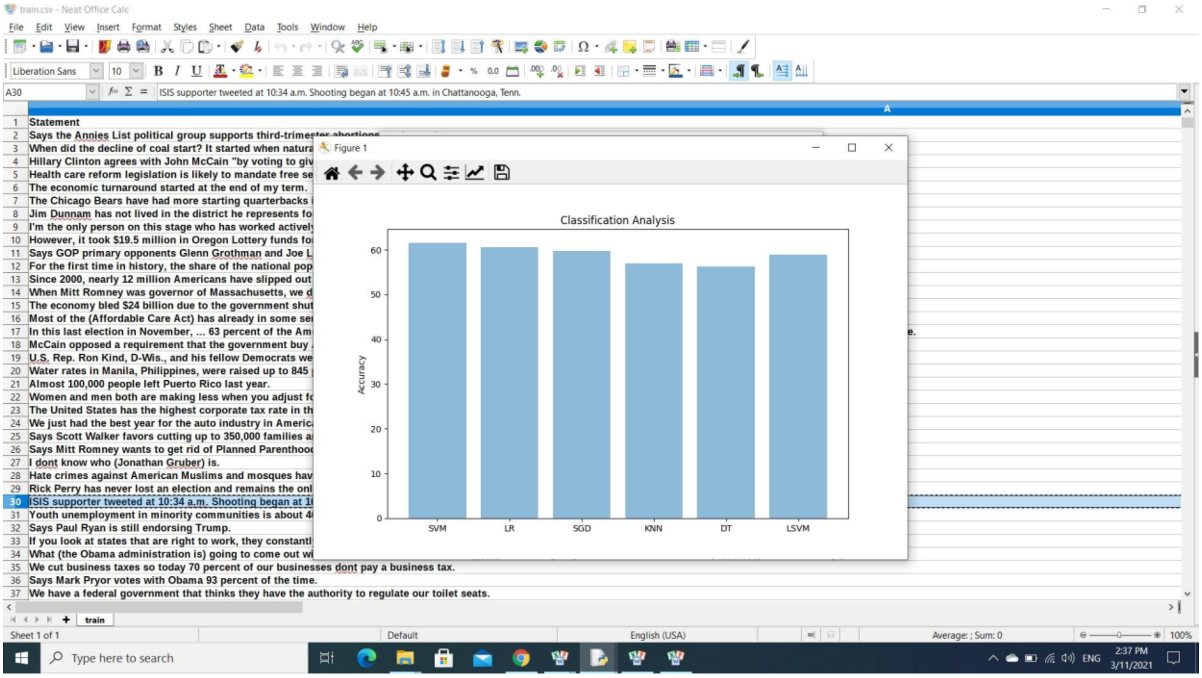
```
1  Statement
2  Says the Annies List political group supports third-trimester abortions
3  When did the decline of coal start? It started when natural gas took off
4  Hillary Clinton agrees with John McCain "by voting to give George Bus
5  Health care reform legislation is likely to mandate free sex change sur
6  The economic turnaround started at the end of my term.
7  The Chicago Bears have had more starting quarterbacks in the last 10
8  Jim Dunnam has not lived in the district he represents for years now.
9  I'm the only person on this stage who has worked actively just last yea
10 However, it took $19.5 million in Oregon Lottery funds for the Port of N
11 Says GOP primary opponents Glenn Grothman and Joe Leibham cast a
12 For the first time in history, the share of the national popular vote mar
13 Since 2000, nearly 12 million Americans have slipped out of the middle
14 When Mitt Romney was governor of Massachusetts, we didnt just slow
15 The economy bled $24 billion due to the government shutdown.
16 Most of the (Affordable Care Act) has already in some sense been waiv
17 In this last election in November, ... 63 percent of the American people          chose not to vote.
18 McCain opposed a requirement that the government buy American-ma
19 U.S. Rep. Ron Kind, D-Wis., and his fellow Democrats went on a spend
20 Water rates in Manila, Philippines, were raised up to 845 percent when
21 Almost 100,000 people left Puerto Rico last year.
22 Women and men both are making less when you adjust for inflation th
23 The United States has the highest corporate tax rate in the free world.
24 We just had the best year for the auto industry in America in history.
25 Says Scott Walker favors cutting up to 350,000 families and children of
26 Says Mitt Romney wants to get rid of Planned Parenthood.
27 I dont know who (Jonathan Gruber) is.
28 Hate crimes against American Muslims and mosques have tripled after
29 Rick Perry has never lost an election and remains the only person to have won the Texas governorship three times in landslide elections.
30 ISIS supporter tweeted at 10:34 a.m. Shooting began at 10:45 a.m. in Chattanooga, Tenn.
31 Youth unemployment in minority communities is about 40 to 45 percent.
32 Says Paul Ryan is still endorsing Trump.
33 If you look at states that are right to work, they constantly do not have budget deficits and they have very good business climates.
34 What (the Obama administration is) going to come out with in the next several months is youre not even going to be able to burn coal very limitedly in the existing plants.
35 We cut business taxes so today 70 percent of our businesses dont pay a business tax.
36 Says Mark Pryor votes with Obama 93 percent of the time.
37 We have a federal government that thinks they have the authority to regulate our toilet seats.
```

Detecting Fake Statements

Load Training Dat: ‹ode/SourceCode/N-Grams/train.csv

Load Testing Dataset: ‹Code/SourceCode/N-Grams/test.csv

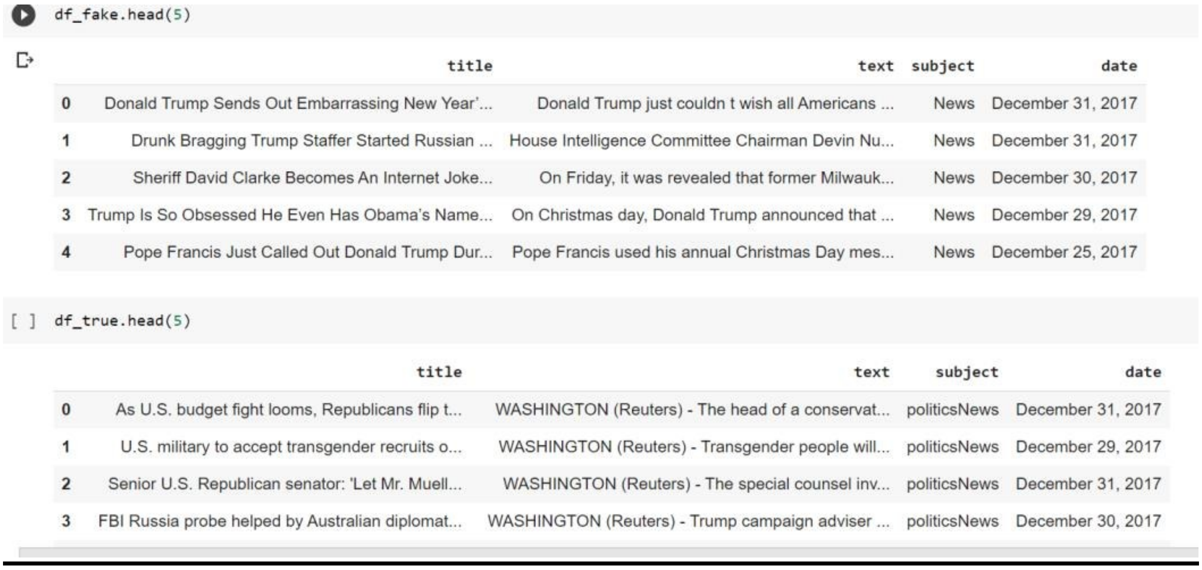
Fig:8.1.1 Detecting Fake News using Dataset



A30 ISIS supporter tweeted at 10:34 a.m. Shooting began at 10:45 a.m. in Chattanooga, Tenn.

	A
1	Statement
2	Says the Annies List political group supports third-trimester abortions
3	When did the decline of coal start? It started when natura
4	Hillary Clinton agrees with John McCain "by voting to giv
5	Health care reform legislation is likely to mandate free se
6	The economic turnaround started at the end of my term.
7	The Chicago Bears have had more starting quarterbacks i
8	Jim Dunnam has not lived in the district he represents fo
9	I'm the only person on this stage who has worked activel
10	However, it took $19.5 million in Oregon Lottery funds fo
11	Says GOP primary opponents Glenn Grothman and Joe L
12	For the first time in history, the share of the national pop
13	Since 2000, nearly 12 million Americans have slipped out
14	When Mitt Romney was governor of Massachusetts, we d
15	The economy bled $24 billion due to the government shut
16	Most of the (Affordable Care Act) has already in some se
17	In this last election in November, ... 63 percent of the Am
18	McCain opposed a requirement that the government buy
19	U.S. Rep. Ron Kind, D-Wis., and his fellow Democrats we
20	Water rates in Manila, Philippines, were raised up to 845
21	Almost 100,000 people left Puerto Rico last year.
22	Women and men both are making less when you adjust fo
23	The United States has the highest corporate tax rate in th
24	We just had the best year for the auto industry in Americ
25	Says Scott Walker favors cutting up to 350,000 families a
26	Says Mitt Romney wants to get rid of Planned Parenthoo
27	I dont know who (Jonathan Gruber) is.
28	Hate crimes against American Muslims and mosques hav
29	Rick Perry has never lost an election and remains the onl
30	ISIS supporter tweeted at 10:34 a.m. Shooting began at 1
31	Youth unemployment in minority communities is about 4
32	Says Paul Ryan is still endorsing Trump.
33	If you look at states that are right to work, they constantl
34	What (the Obama administration is) going to come out wi
35	We cut business taxes so today 70 percent of our businesses dont pay a business tax.
36	Says Mark Pryor votes with Obama 93 percent of the time.
37	We have a federal government that thinks they have the authority to regulate our toilet seats.

Fig:8.1.2 Accuracy level of Algorithms with dataset



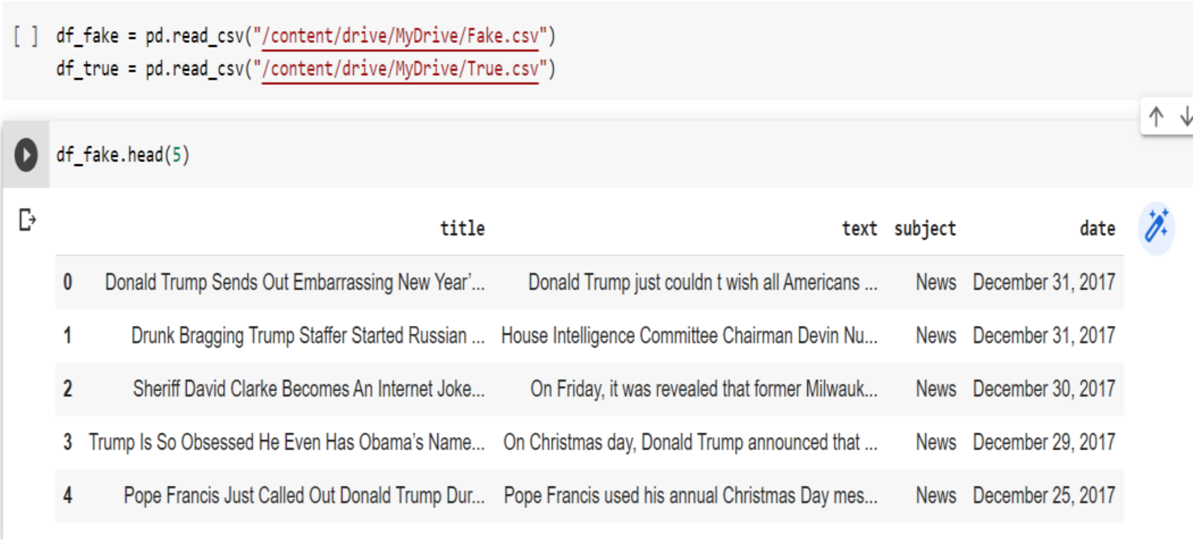
```
df_fake.head(5)
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

```
[ ] df_true.head(5)
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017

Fig.7.2 : Fake.csv and True.csv

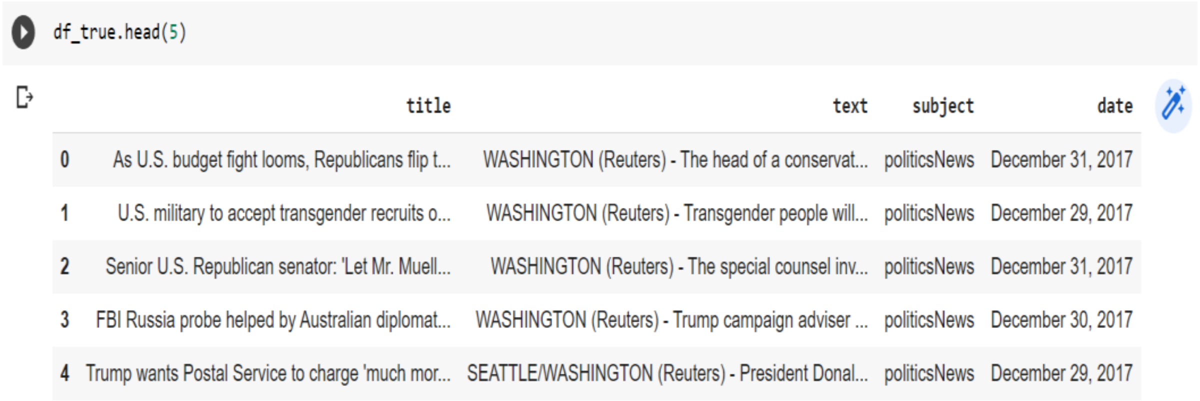


```
[ ] df_fake = pd.read_csv("/content/drive/MyDrive/Fake.csv")
    df_true = pd.read_csv("/content/drive/MyDrive/True.csv")
```

```
df_fake.head(5)
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

Fig.7.3 :fake.csv

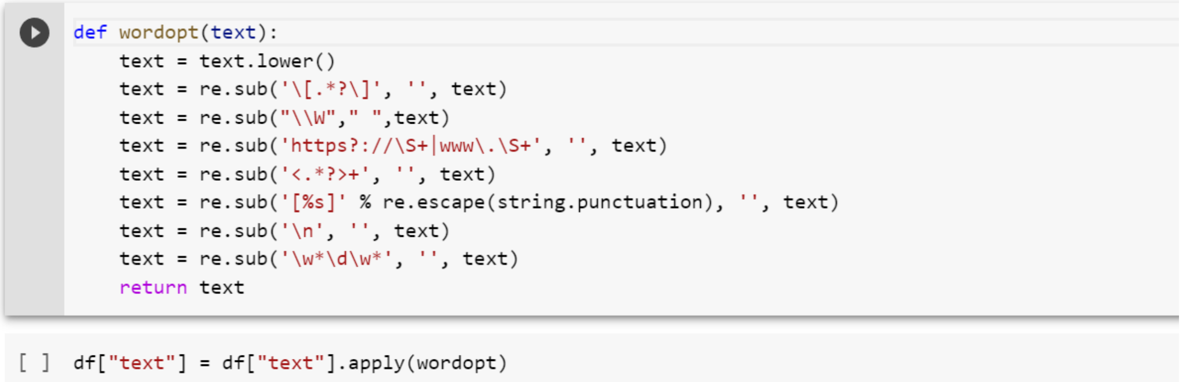


```
df_true.head(5)
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

Fig.7.4 :true.csv

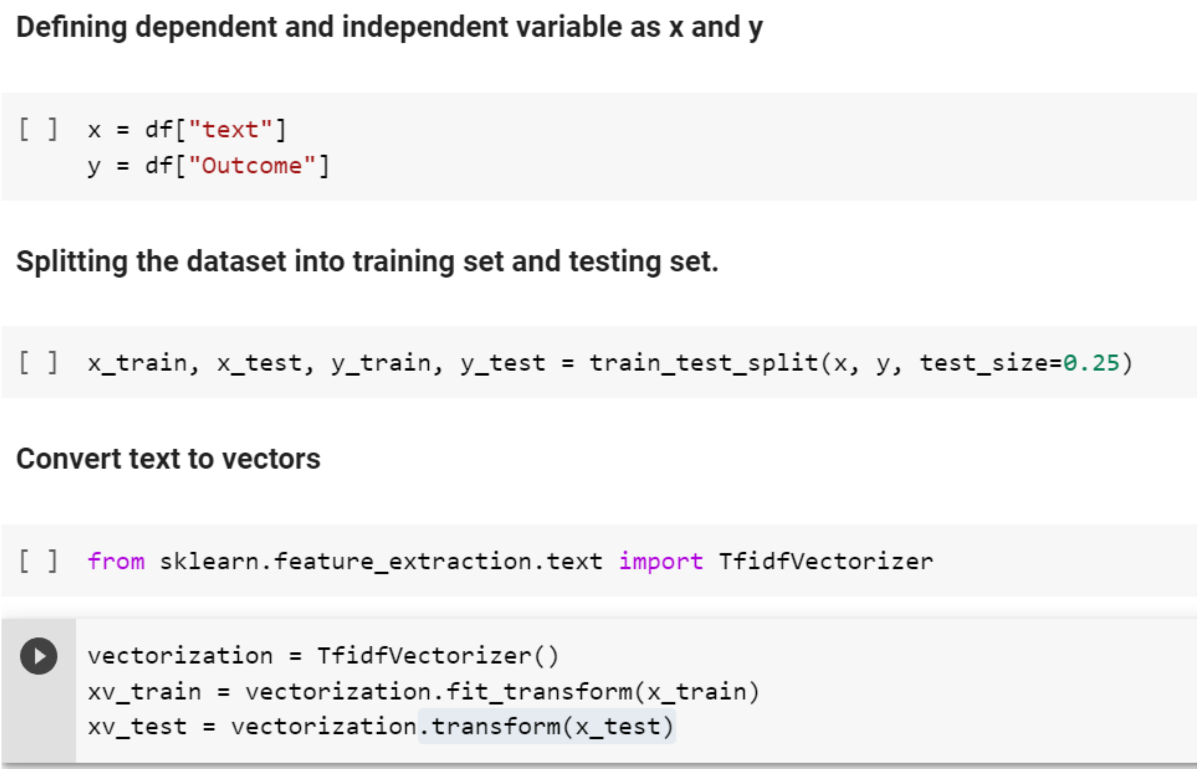
Creating a function to convert the text in lowercase, remove the extra space, special chr., ulr and links



```
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
[ ] df["text"] = df["text"].apply(wordopt)
```

Fig. 7.5: Pre-processing task of words



Defining dependent and independent variable as x and y

```
[ ]  x = df["text"]
     y = df["Outcome"]
```

Splitting the dataset into training set and testing set.

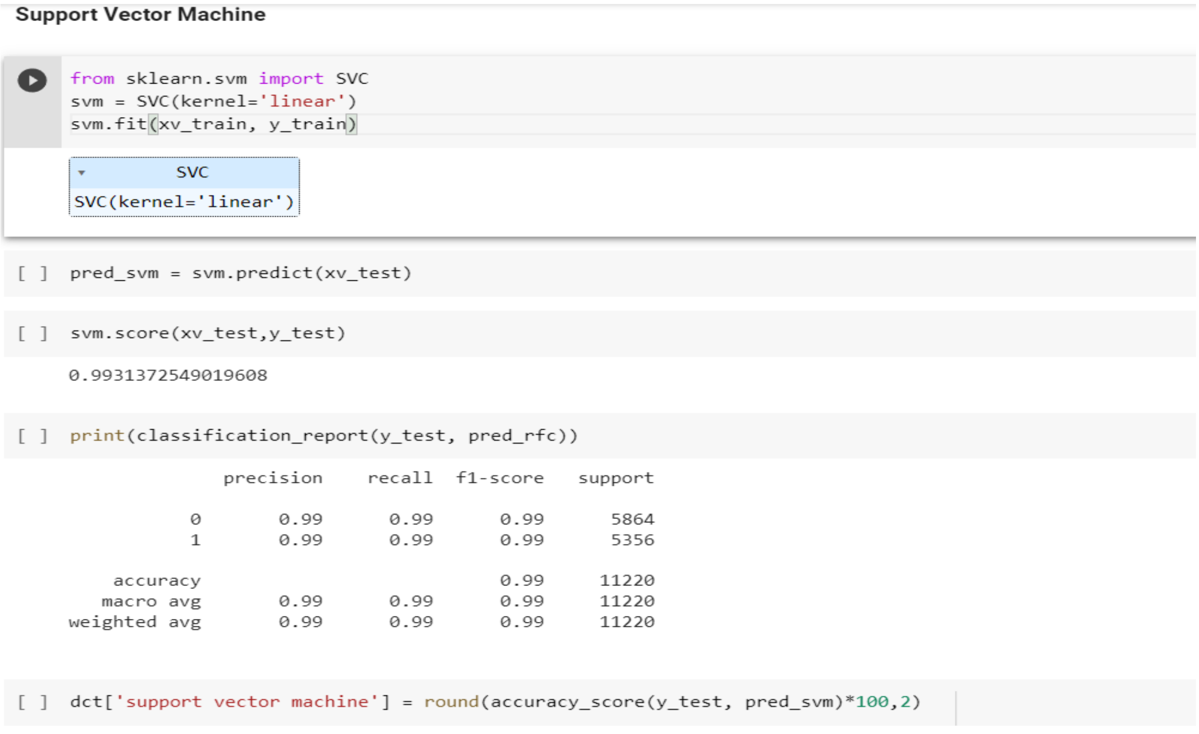
```
[ ]  x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Convert text to vectors

```
[ ]  from sklearn.feature_extraction.text import TfidfVectorizer
```

```
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)
```

Fig. :7.6 Train-Test Split



Support Vector Machine

```
from sklearn.svm import SVC
svm = SVC(kernel='linear')
svm.fit(xv_train, y_train)
```

```
  SVC
SVC(kernel='linear')
```

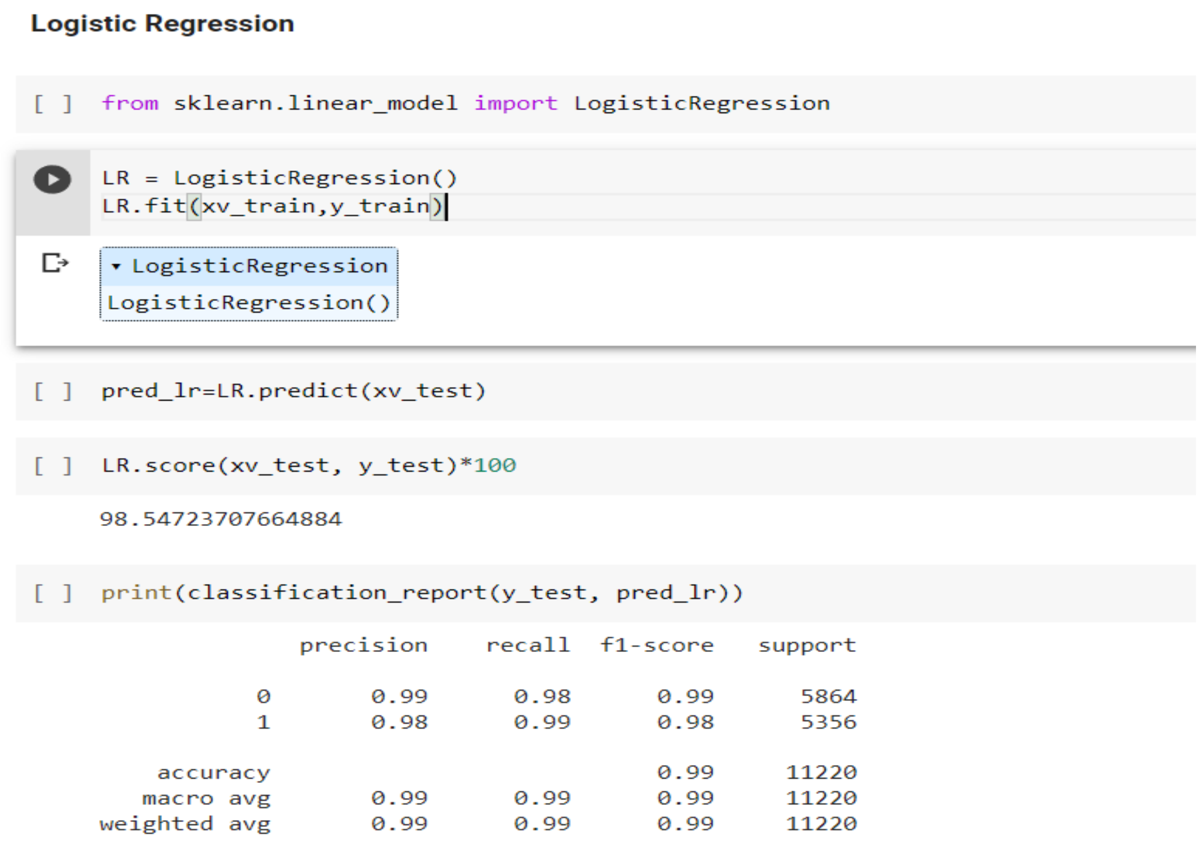
```
[ ]  pred_svm = svm.predict(xv_test)
```

```
[ ]  svm.score(xv_test,y_test)

     0.9931372549019608
```

```
[ ]  print(classification_report(y_test, pred_rfc))

                   precision    recall  f1-score   support

                0       0.99      0.99      0.99      5864
                1       0.99      0.99      0.99      5356

         accuracy                           0.99     11220
        macro avg       0.99      0.99      0.99     11220
     weighted avg       0.99      0.99      0.99     11220
```

```
[ ]  dct['support vector machine'] = round(accuracy_score(y_test, pred_svm)*100,2)
```

Fig.8.3.2 SVM



Logistic Regression

```
from sklearn.linear_model import LogisticRegression
```

```
LR = LogisticRegression()
LR.fit(xv_train,y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

```
pred_lr=LR.predict(xv_test)
```

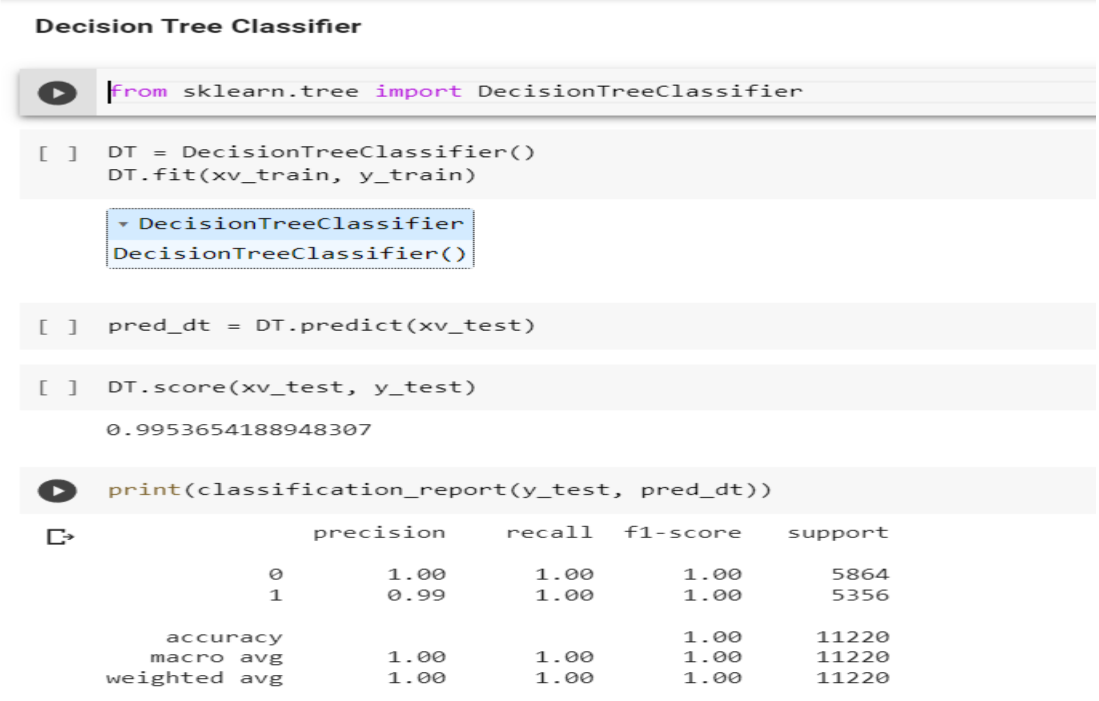
```
LR.score(xv_test, y_test)*100
```

```
98.54723707664884
```

```
print(classification_report(y_test, pred_lr))
```

```
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      5864
           1       0.98      0.99      0.98      5356

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Fig.8.3.3 logistic regression



Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier
```

```
DT = DecisionTreeClassifier()
DT.fit(xv_train, y_train)
```

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

```
pred_dt = DT.predict(xv_test)
```

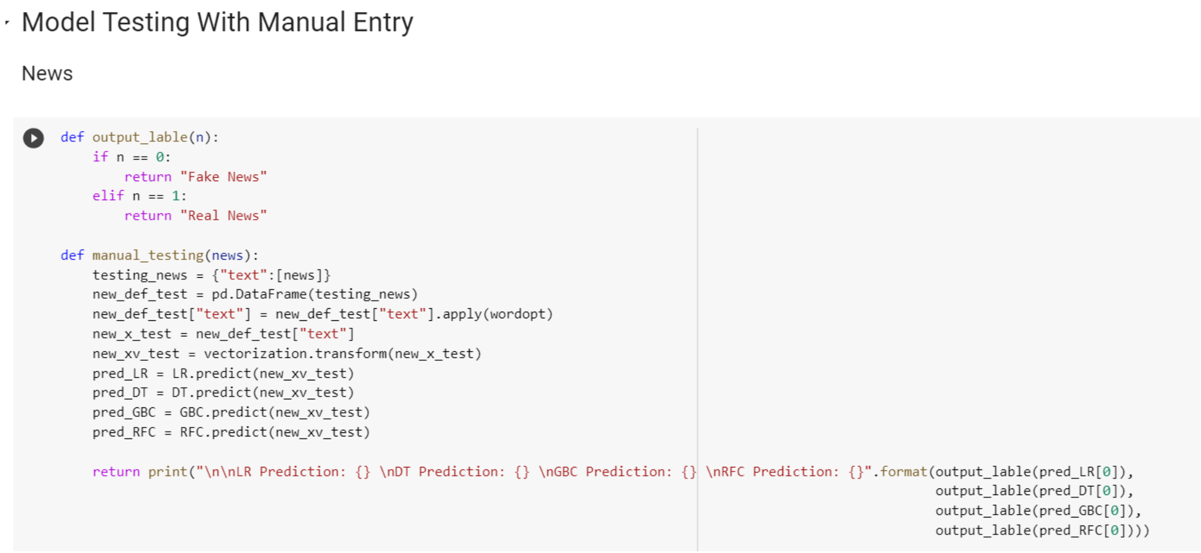
```
DT.score(xv_test, y_test)
```

```
0.9953654188948307
```

```
print(classification_report(y_test, pred_dt))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5864
           1       0.99      1.00      1.00      5356

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Fig.8.3.4

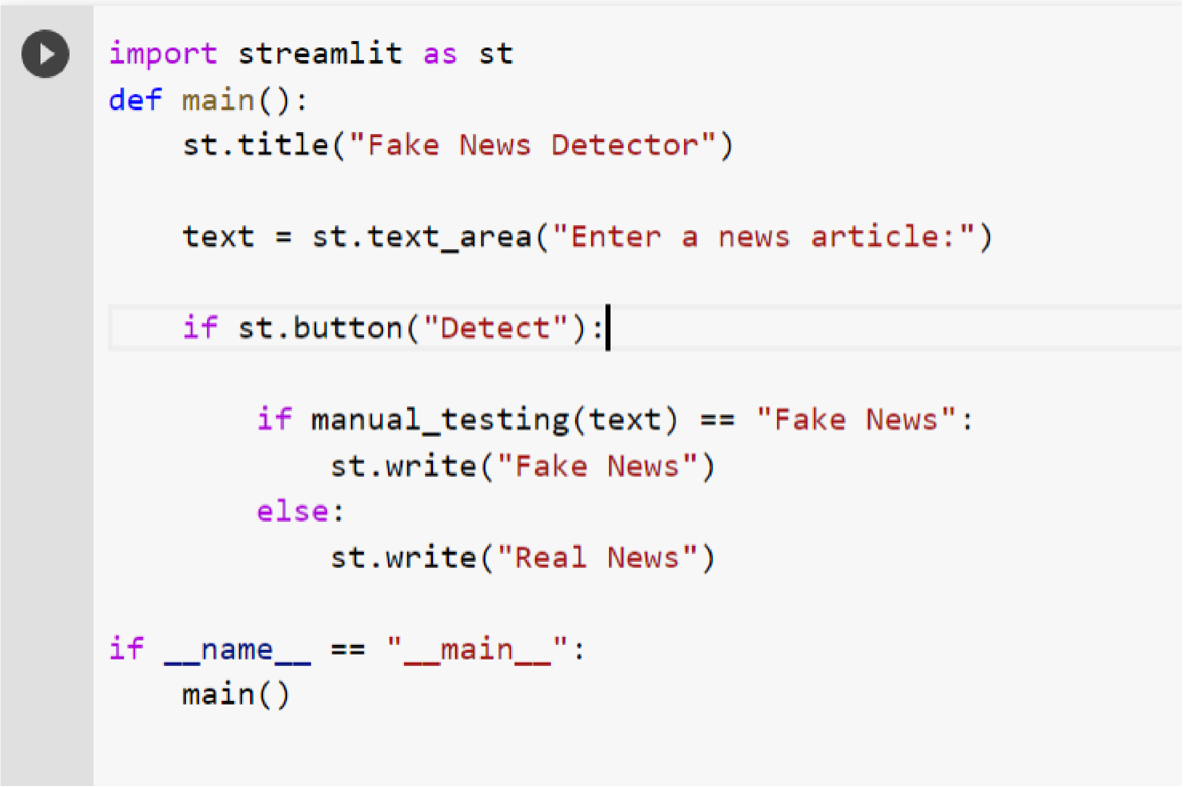


Model Testing With Manual Entry

News

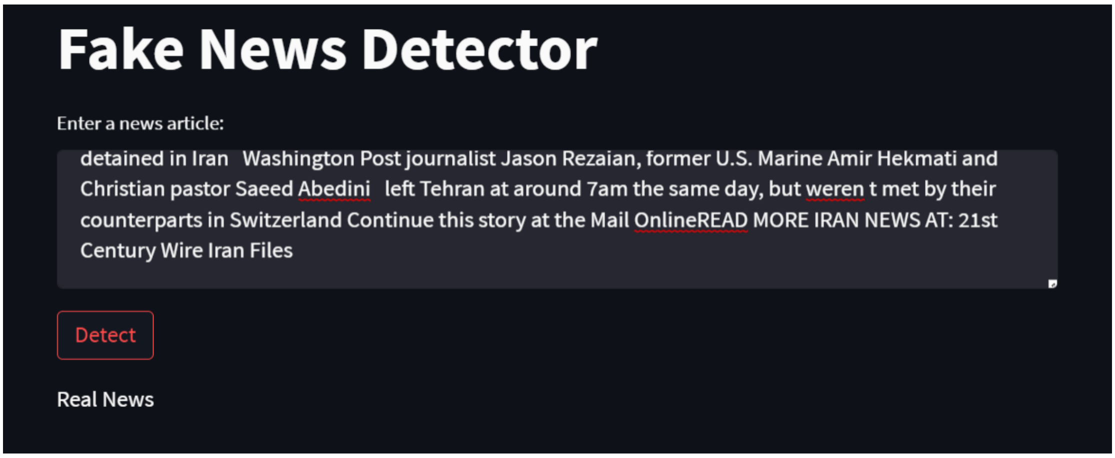
```
def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Real News"

def manual_testing(news):
    testing_news = {"text":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xv_test = vectorization.transform(new_x_test)
    pred_LR = LR.predict(new_xv_test)
    pred_DT = DT.predict(new_xv_test)
    pred_GBC = GBC.predict(new_xv_test)
    pred_RFC = RFC.predict(new_xv_test)

    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC Prediction: {} \nRFC Prediction: {}".format(output_lable(pred_LR[0]),
                                                                                                              output_lable(pred_DT[0]),
                                                                                                              output_lable(pred_GBC[0]),
                                                                                                              output_lable(pred_RFC[0])))
```

Fig.7.7 TESTING



```
import streamlit as st
def main():
    st.title("Fake News Detector")

    text = st.text_area("Enter a news article:")

    if st.button("Detect"):

        if manual_testing(text) == "Fake News":
            st.write("Fake News")
        else:
            st.write("Real News")

if __name__ == "__main__":
    main()
```

FIG.7.8 :SAMPLE INPUT



Fake News Detector

Enter a news article:

detained in Iran   Washington Post journalist Jason Rezaian, former U.S. Marine Amir Hekmati and Christian pastor Saeed Abedini   left Tehran at around 7am the same day, but weren t met by their counterparts in Switzerland Continue this story at the Mail OnlineREAD MORE IRAN NEWS AT: 21st Century Wire Iran Files

Real News

FIG.7.9 : OUTPUT

CHAPTER 8

CODING

8. SOURCE CODE

8.1 ADMIN PAGE

```
from PyQt5 import QtCore, QtGui, QtWidgets
from Admin import Ui_Admin
import pandas as pd
class Ui_Dialog(object):
def admin(self, event):
try:
self.admn = QtWidgets.QDialog()
self.ui = Ui_Admin(self.admn)
self.ui.setupUi(self.admn)
self.admn.show()
except Exception as e:
print(e.args[0])
tb = sys.exc_info()[2]
print(tb.tb_lineno)
event.accept()
def setupUi(self, Dialog):
Dialog.setObjectName("Dialog")
Dialog.resize(702, 435)
Dialog.setStyleSheet("background-color: rgb(0, 85, 127);")
self.label = QtWidgets.QLabel(Dialog)
self.label.setGeometry(QtCore.QRect(60, 60, 601, 41))
self.label.setStyleSheet("color: rgb(255, 255, 255);\n"
"font: 75 18pt \"Tahoma\";")
```

```
self.label.setObjectName("label")
self.label_2 = QtWidgets.QLabel(Dialog)
self.label_2.setGeometry(QtCore.QRect(200, 150, 261, 181))35
self.label_2.setStyleSheet("image: url(../N-Grams/images/admin.png);")
self.label_2.setText("")
self.label_2.setObjectName("label_2")
self.label_2.mousePressEvent = self.admin
self.retranslateUi(Dialog)
QtCore.QMetaObject.connectSlotsByName(Dialog)
def retranslateUi(self, Dialog):
_translate = QtCore.QCoreApplication.translate
Dialog.setWindowTitle(_translate("Dialog", "Online Fake News"))
self.label.setText(_translate("Dialog", "Detection of Online Fake News Using N
Gram Analysis"))
if _name_ == "_main_":
import sys
app = QtWidgets.QApplication(sys.argv)
Dialog = QtWidgets.QDialog()
ui = Ui_Dialog()
ui.setupUi(Dialog)
Dialog.show()
sys.exit(app.exec_())
```

8.1 following script as login.html

```
<html>
<body>
<formaction="http://localhost:5000/login"method="post">
<p>Enter Name:</p>
<p><inputtype="text"name="nm"/></p>
<p><inputtype="submit"value="submit"/></p>
</form>
```

```
</body>
</html>
```

Now enter the following script in Python shell.

```
from flask importFlask, redirect,url_for, request
app=Flask( name )
@app.route('/success/<name>')
def success(name):
return'welcome %s'% name
@app.route('/login',methods=['POST','GET'])
def login():
ifrequest.method=='POST':
user=request.form['nm']
return redirect(url_for('success',name= user))
else:
user=request.args.get('nm')
return redirect(url_for('success',name= user))
if
name ==' main ':
app.run(debug =True)
```

After the development server starts running, open login.html in the browser, enter name in the text field and click Submit

8.3Algorithms

8.3.1 Naive Bayes

- One of supervised learning algorithm based on probabilistic classification technique.
- It is a powerful and fast algorithm for predictive modelling.
- In this project, I have used the Multinomial Naive Bayes Classifier.

8.3.2.Support Vector Machine- SVM

- SVM‟s are a set of supervised learning methods used for classification, and regression.
- Effective in high dimensional spaces.
- Uses a subset of training points in the support vector, so it is also memory Efficient.

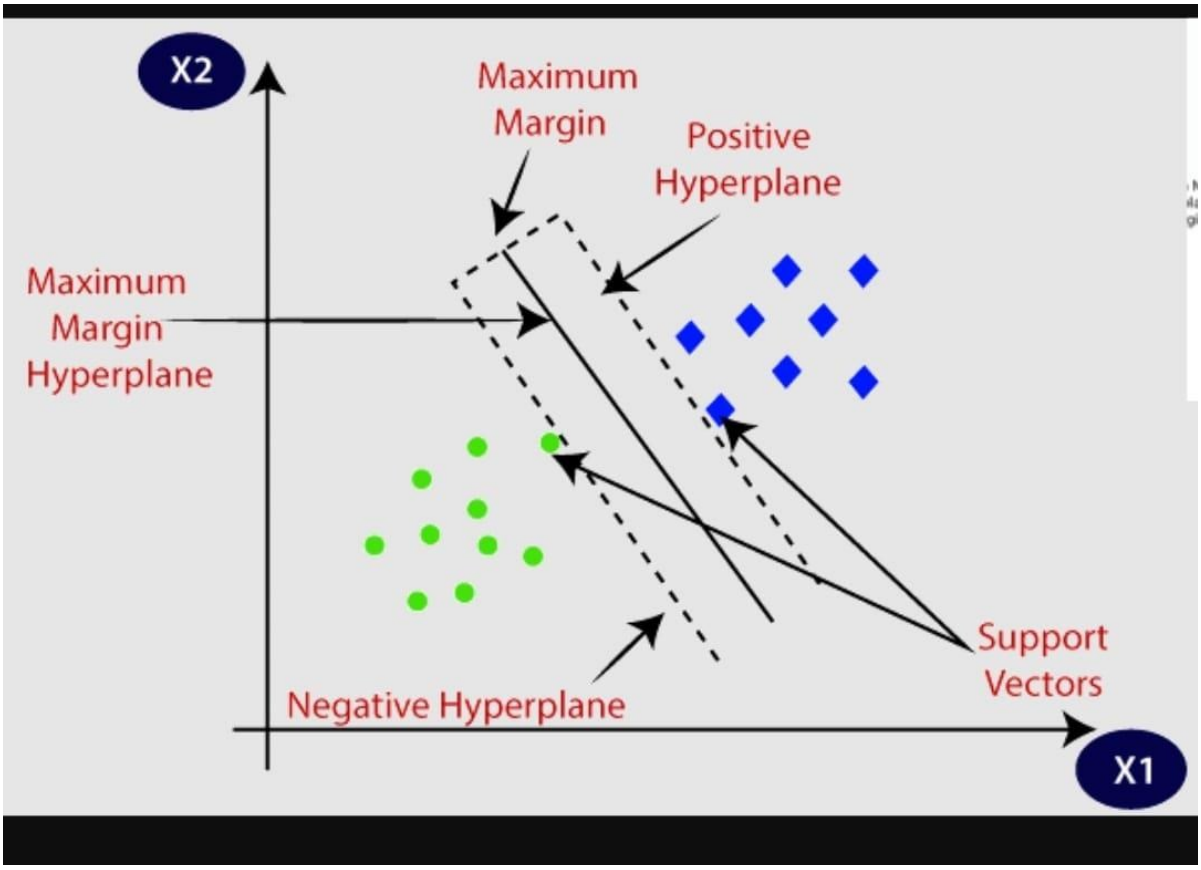
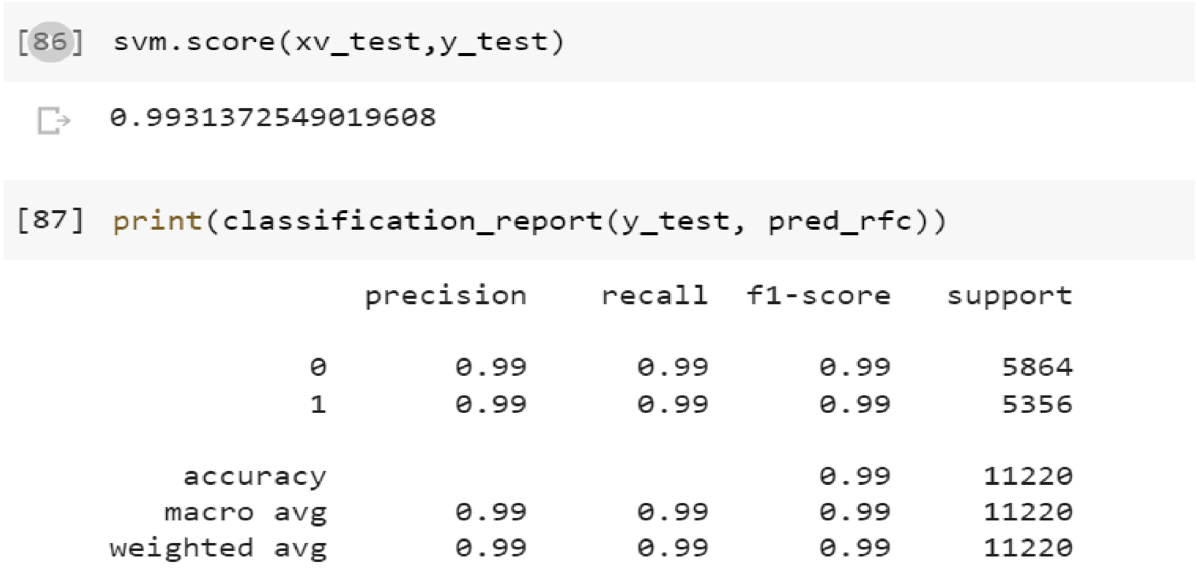


Fig.8.3.2

Below are the Results from applying Support Vector Machine model:

Table 1: Classification Report of SVM



```
[86] svm.score(xv_test,y_test)

     0.9931372549019608
```

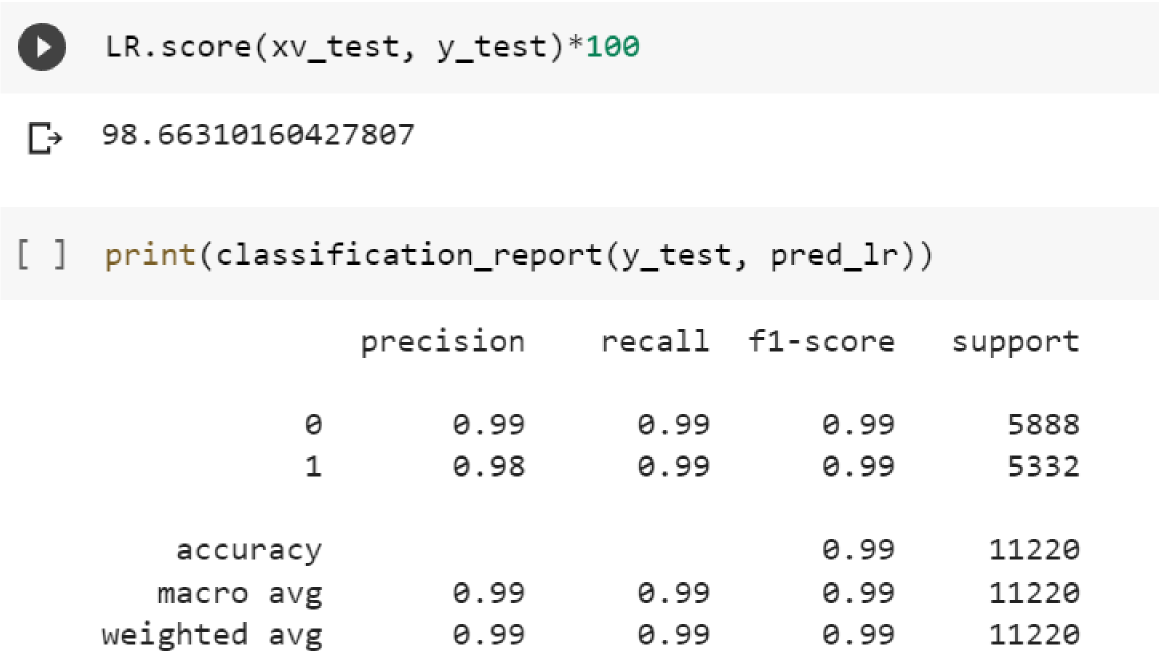
```
[87] print(classification_report(y_test, pred_rfc))

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5864
           1       0.99      0.99      0.99      5356

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

8.3.3Logistic Regression

- Linear model for classification rather than regression.
- The expected values of the response variable are modeled based on combination of values taken by the predictors

Below are the Results from applying Logistic Regression model:

Table 2: Classification Report of LR

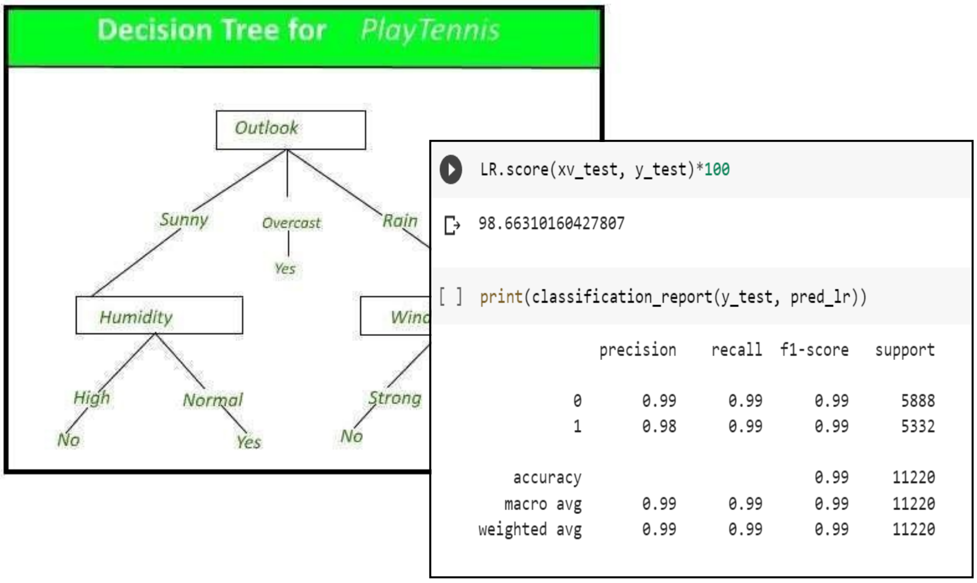


```
LR.score(xv_test, y_test)*100

98.66310160427807

[ ] print(classification_report(y_test, pred_lr))

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5888
           1       0.98      0.99      0.99      5332

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

8.3.4 Decision Tree Classification

▪ For both binary and multi-class classification tasks, decision tree classification is a popular machine learning approach. The input data are recursively divided into subgroups depending on the most instructive characteristic.

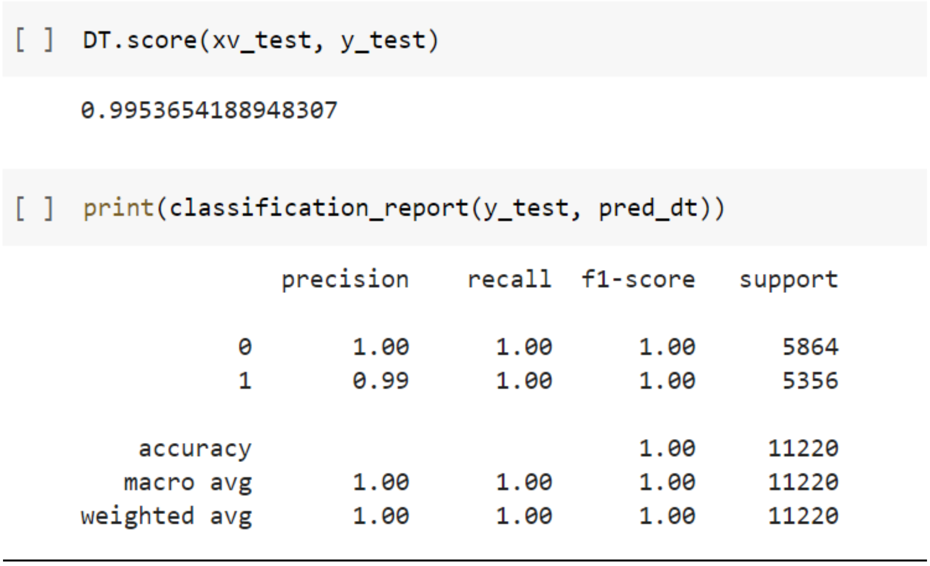
▪ decision trees can handle category and numerical data and are simple to understand and use. Additionally, they are resistant to noise and missing data and are capable of capturing intricate non-linear correlations between features.



```
LR.score(xv_test, y_test)*100

98.66310160427807

[ ] print(classification_report(y_test, pred_lr))

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5888
           1       0.98      0.99      0.99      5332

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Below are the Results from applying Decision Tree Classification model:

Table 3: Classification Report from Decision Tree



```
[ ]  DT.score(xv_test, y_test)

     0.9953654188948307
```

```
[ ]  print(classification_report(y_test, pred_dt))

                   precision    recall  f1-score   support

                0       1.00      1.00      1.00      5864
                1       0.99      1.00      1.00      5356

         accuracy                           1.00     11220
        macro avg       1.00      1.00      1.00     11220
     weighted avg       1.00      1.00      1.00     11220
```

8.3.5 Natural Language Processing (NLP)

The study of how computers interact with human (natural) languages is known as natural language processing, or NLP, and it is a branch of computer science and artificial intelligence that focuses on instructing computers to efficiently analyse massive volumes of natural language data. In the fields of linguistics, computer science, information engineering, and artificial intelligence, natural language processing (NLP) studies how computers interact with human (natural) languages. Its major goal is to instruct computer programmers in how to study and analyses vast amounts of natural language.

CHAPTER 9

PROJECT LIMITATION AND FUTURE SCOPE

9. PROJECT LIMITATION

1. **Algorithmic Limitations**: Fake news detection algorithms rely heavily on pattern recognition, natural language processing, and machine learning techniques. However, these algorithms may not always be accurate or reliable, especially when dealing with subtle forms of misinformation or evolving tactics employed by purveyors of fake news.
2. **Bias in Data**: Fake news detection models are trained on datasets that may themselves be biased or incomplete. This can lead to algorithmic biases that affect the accuracy of the detection process, particularly if the training data is not representative of the full spectrum of fake news content.
3. **Rapidly Evolving Tactics**: Those who create and spread fake news are often quick to adapt to detection methods, employing new tactics and strategies to evade detection. As a result, fake news detection systems may struggle to keep pace with these changes, leading to a constant game of cat and mouse between detectors and purveyors of misinformation.
4. **Contextual Understanding**: Fake news detection algorithms may struggle to understand the nuances of language and context, leading to false positives or negatives. Some fake news stories may be based on a kernel of truth or may use ambiguous language to evade detection, making it difficult for algorithms to accurately classify them.
5. **Privacy Concerns**: Some effective methods of detecting fake news, such as tracking the spread of information across social networks or analyzing user behavior, raise significant privacy concerns. Balancing the need for effective detection with the protection of user privacy is a complex challenge.

6. **Legal and Ethical Considerations**: Determining what constitutes fake news can be subjective and may vary depending on cultural, political, and social factors. Moreover, there are legal and ethical considerations surrounding the censorship of information, even if it is deemed to be fake. This creates a dilemma for platforms and governments seeking to combat fake news while upholding freedom of speech and expression.
7. **Human Factors**: Fake news detection often requires human judgment to validate algorithmic results or to make nuanced decisions that algorithms struggle with. However, human reviewers may introduce their own biases or errors into the process, undermining the effectiveness of the detection system.

9.1 FUTURE SCOPE

Future research and advancement in the field of false news detection area bundantly possible. Future efforts to identify bogus news may go in the following directions:

➢ **Including more varied and subtle aspects:** For the most part, current methods for detecting false news rely on simple text-based traits like TF-IDF vectors or bag-of-words. Research in the future could concentrate on more complex and diverse aspects, such sentiment analysis, network analysis, or multimedia analysis (for instance, identifying false images or videos).

➢ **Creating more interpretable models:** Existing methods for spotting fake news sometimes rely on complex machine learning algorithms that might be difficult to comprehend. In the future, it would be beneficial to develop more intelligible models that might provide more information on how people make decisions.

➢ **Combining information from other sources:** In addition to social media, news articles, and videos, fake news is regularly spread through other media channels and platforms. The development of methods that can incorporate data from several sources may be crucial in the future to improve false news identification.

- **Adapting to shifting strategies:** It will be crucial for fake news detection technologies to develop alongside the tactics used by those who create and spread it. For this, the detection methods might need to be regularly reviewed and improved.

CHAPTER 10

REFERENCES

https://www.google.com/

https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/

https://paperswithcode.com/task/fake-news-detection

https://www.kaggle.com/datasets/sumanthvrao/fakenewsdataset