**FAKE NEWS DETECTION**

**USING PYTHON**

***A***

***Major Project Report Submitted***

***In partial fulfilment***

***For the award of the Degree of***

**BACHELOR OF TECHNOLOGY**

***In Department of Computer science and Engineering***

|  |  |  |
| --- | --- | --- |
| **Submitted By:** | **Guided By:** | **Submitted To:** |
| Diksha Bapna | Mr. Aaditya Maheshwari | HOD CS Department |
|  | (Project In charge) |  |
|  |  |  |
|  |  |  |



#### Department of Computer Science and Engineering TECHNO INDIA NJR INSTITUTE OF TECHNOLOGY RAJASTHAN TECHNICAL UNIVERSITY

**May 2023**



**Department of Computer Science and Engineering**

#### Techno India NJR Institute of Technology, Udaipur

**CERTIFICATE**

This is to certify that project work titled **“Fake News Detection using Python”** by **“Diksha Bapna”** was successfully carried out in the **Department of Computer Science and Engineering, Techno India NJR Institute of Technology**, and the report is approved for submission in partial fulfillment of the requirements for the award of degree of Bachelors of Technology in Computer Science and Engineering.

#### Aaditya Maheshwari

**Assistant Professor, Project In charge**

Department of Computer Science



Department of Computer Science and Engineering

Techno India NJR Institute of Technology, Udaipur-313001

**Examiner Certificate**

This is to certify that the following student **Diksha Bapna** of final year B.Tech. (Computer Science and Engineering), was examined for the project work titled **Fake News Detection Using Python** during the academic year 2023 at Techno India NJR Institute of Technology, Udaipur.

**Remarks:**

**Date:**

Signature Signature

(**Internal Examiner**) (**External Examiner**)

Name: - ……………………… Name: - ………………………

Designation:- ……………….. Designation:- ………………..

Department: - ………………. Department: - ……………….

Organization:- ……………… Organization:- ………………

## PREFACE

In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by these online fake news easily, which has brought about tremendous effects on the offline society already. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely. This report aims at investigating the principles, methodologies and algorithms for detecting fake news articles, creators and subjects from online social networks and evaluating the corresponding performance. Information preciseness on Internet, especially on social media, is an increasingly important concern, but web-scale data hampers, ability to identify, evaluate and correct such data, or so called "fake news," present in these platforms. In this paper, we propose a method for "fake news" detection and ways to apply it on Facebook, one of the most popular online social media platforms. This method uses Naive Bayes classification model to predict whether a post on Facebook will be labeled as real or fake. The results may be improved by applying several techniques that are discussed in the report. Received results suggest, that fake news detection problem can be addressed with machine learning methods.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

TECHNO INDIA NJR INSTITUTE OF TECHNOLOGY, UDAIPUR-313001

# ACKNOWLEDGEMENT

It gives me immense pleasure to express my sense of gratitude and sincere thanks to my respected and esteemed guide **Mr. Aaditya Maheshwari (Assistant Professor, Project In charge), TINJRIT** for their valuable guidance, encouragement and help for completing this work.

I would like to express my sincere thanks to **Head of Department, Dept. of CSE TINJRIT** for giving me this opportunity to undertake this project.

I also wish to express my indebtedness to my parents as well as my family member whose blessings and support always helped me to face the challenges ahead.

|  |  |
| --- | --- |
| **Place: Udaipur** | **Diksha Bapna**  **(19ETCCS014)** |

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

TECHNO INDIA NJR INSTITUTE OF TECHNOLOGY, UDAIPUR-313001

## 

## TABLE OF CONTENT

[Preface. iii](#_bookmark0)

[Acknowledgement iv](#_bookmark1)

[Table Of Content v](#_bookmark2)i

[List Of Figures vii](#_bookmark3)

CHAPTER – 1 Introduction 1-4

* 1. [Overview 2](#_bookmark6)
  2. Problem Statement 2
  3. [Learning Techniques 3](#_bookmark7)
  4. [Flow Chart 3](#_bookmark8)
  5. Python Libraries and Algorithms 3
  6. [Addressing ethical and social issues and responsibilities 4](#_bookmark9)

CHAPTER – 2 Literature Survey 5-10

* 1. [Media Rich Fake news detection: A Survey 6](#_bookmark11)
  2. [Fake News Detection in Social media 7](#_bookmark12)
  3. [The Spread of Fake news by Social Bots 8](#_bookmark13)
  4. [Misleading Online Content 8](#_bookmark14)

CHAPTER – 3 Methodology 11-16

* 1. [The Dataset 12](#_bookmark17)
  2. [Programming Language and Machine Learning Model 14](#_bookmark18)
  3. [Cost Benefit analysis 15](#_bookmark19)
  4. [Use-Case Diagram 16](#_bookmark20)

CHAPTER – 4 Results and Discussion 17-19

* 1. [Requirement Analysis 18](#_bookmark21)

4.1.1 Functional Requirements 19

4.1.2 Non Functional Requirements 20

* 1. Testing [19](#_bookmark22)

CHAPTER – 5 Conclusion And Future Enhancements 36-37

* 1. [Conclusion …3](#_bookmark30)7

5.2 [Future Scope](#_bookmark31) [3](#_bookmark31)7

**CHAPTER - 6 Appendix** 38-39

## LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Name** | **Page No.** |
| Figure 1 | Flow Chart Diagram | 7 |
| Figure 2 | Naïve Bayes Algorithm | 12 |
| Figure 3 | Dataset Used | 13 |

# CHAPTER – 1

# INTRODUCTION

## INTRODUCTION

#### Overview

#### These days, fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is “fake news” but lately blathering social media’s discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. The importance of disinformation within American political discourse was the subject of weighty attention, particularly following the American president election. The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper, it is seemed to produce a model that can accurately predict the likelihood that a given article is fake news. Facebook has been at the epicenter of much critique following media attention. They have already implemented a feature to flag fake news on the site when a user sees’ it; they have also said publicly they are working on to distinguish these articles in an automated way. Certainly, it is not an easy task. A given algorithm must be politically unbiased – since fake news exists on both ends of the spectrum – and also give equal balance to legitimate news sources on either end of the spectrum. In addition, the question of legitimacy is a difficult one. However, in order to solve this problem, it is necessary to have an understanding on what Fake News is.

#### Problem Statement

#### In this day and age, it is extremely difficult to decide whether the news we come across is real or not. There are very few options to check the authenticity and all of them are sophisticated and not accessible to the average person. There is an acute need for a web-based fact-checking platform that harnesses the power of Machine Learning to provide us with that opportunity.

#### Social media have enhanced the experience of news consumption due to its cost effective, easily accessible and widely distributable characteristic. However, it has made an average internet user easily vulnerable to consuming news that is intentionally or unintentionally distorted which can have drastic consequences and puts an individual and society at risk.

#### Through this project, we aim to obtain maximum accuracy in fake news detection and real news generation to obtain a perfect result.

#### 1.3 Learning Techniques

#### We have used various algorithms and techniques to urge the specified results. Machine learning algorithms such as passive aggressive classifier and Naïve Bayes algorithms are used to predict the output and with a good accuracy. Firstly, data pre-processing is done with stemming and stop words. This process helps in cleaning up the data. After the pre-processing, feature extraction takes place. This is achieved by TFIDF vectorizer. The TFIDF will check how significant a word is in the whole document. Thus, after the machine learning algorithms the news is predicted to be real or fake. If the news is found out to be fake, the data is extracted and we have used Keyword search algorithm which is achieved by using RAKE and thus we will be able to get the output i.e., related news from some of the trusted sites.

#### 1.4 Flow Chart

#### a) Data is a Pre-processing-In this model, we take input and preprocess the data using stemming and stop words. Thus, the data is cleaned in this model.

#### b) Feature Extraction- Using TFIDF, how significant a word is in the whole document is checked.

#### c) Classifier Training- After training and testing the data, ML algorithms like passive aggressive classifier and Naïve Bayes algorithm is used to predict the value.

#### d) Data extraction- After the predicted output, Keyword extraction algorithm is used, which is implemented using rake.

#### 

#### Figure 1

**1.5 Python Libraries and Algorithms**

The various Machine Learning algorithms that we used are deployed in python programming language which utilizes a variety of open-source python libraries.

1. Libraries and algorithms

• NumPy (numerical python, integrates C/C++) - NumPy is the elementary package of the Python language for scientific computing.

• Pandas (open source library, data manipulation) - Pandas library is made upon NumPy, that means Pandas wants Numpy to control. Pandas offers a technique to produce, manipulate and altercate the data.

• Sklearn (open source ML library, supervised and unsupervised learning algorithm) - It offers numerous algorithms like the ones which can support random forests, vector machine, and k-neighbours, and with all this it also helps to support the Python numerical and scientific libraries like SciPy and NumPy.

• Scikit (free ML library, features SVM, classification and regression) - This library is absorbed on modeling of data. Scikit does not emphasis on loading the data, manipulation and summarization of data.

• Passive Aggressive Classifier: - Passive Aggressive Classifiers are an online learning algorithm family that functions in the same way as a perceptron since they do not need a learning rate. Such a classifier remains passive when the classification outcome is correct, however it turns aggressive as soon as it comes across an incorrect outcome in the event of a miscalculation, after which it updates and modifies the unwanted outcome. In this project, such a classifier can help detect fake news and then fetch and generate relevant, genuine news to the user in the process from trusted news sources, thus fulfilling its purpose of making the 17 much-needed modifications that corrects the loss. Due to its simplicity in terms of implementation as well as its quality to be used for incremental large-scale learning, it plays an imperative role in classifier training stage after a dataset has been through a test-train split procedure in order to estimate and enhance the performance of the machine learning model used in this project.

• Naïve Bayes Algorithm: - Naïve Bayes Algorithm is a family of classification algorithms which works on the principle of Bayes Theorem. Therefore, it is also known as a collection of probabilistic classifiers and can be implemented in various classification tasks. In such an algorithm, all pairs of features which are classified are independent of each other. Some of its applications include filtering spam, sentiment prediction and classification of documents. Naïve Bayes holds great significance in this project when it comes to classifying a news article as real news or fake news since it is highly scalable, efficient and can be used to produce real-time predictions while handling continuous as well as discrete data.

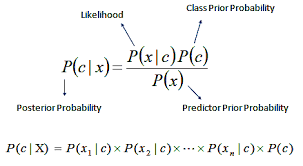
****

Figure 2

* **Keyword Search Algorithm**: - Keyword Search Algorithms is a text analysis technique which can be used to determine key phrases in a text in order to simplify information extraction. In this python project, feature extraction methods such as TF-IDF (Term Frequency – Inverse Document Frequency) method has been implemented which makes use of numerical statistic in order to assign a weighting factor based on the frequency of a word in a collection of documents which can determine its importance for information retrieval.

**1.6 ADDRESSING ETHICAL AND SOCIAL ISSUES AND RESPONSIBILITIES**

Our ethical and social responsibilities include: -

• Verification of the genuineness of the trusted-sources from where the fake news will be detected.

• The generalized models proposed are not going to be in favour of any political, social or economic organization.

• We respect the copyrights, acknowledge the contributions to our research.

• We are always open to fresh thoughts and critique.

• As social media users, encouraging everyone to play their part of personal responsibility of double-checking the information they consume instead of demanding social media companies/journalists to play the role.

# CHAPTER – 2

**LITERATURE SURVEY**

## LITERATURE SURVEY

The available literatu**re** has described many automaticdetection techniques of fake news and deception posts. Since there are multidimensional aspects of fake news detection ranging from using chat bots for spread of misinformation to use of click baits for the rumor spreading. There are many click baits 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.

**2.1 MEDIA RICH FAKE NEWS DETECTION: A SURVEY**

In general, the goal is profiting through click baits. Click baits 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.

**2.2 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. In general, the goal is profiting through click baits. Click baits 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.

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 recent years, the reliability of information on the Internet has emerged as a crucial issue of modern society. Social network sites 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 THE SPREAD OF FAKE NEWS BY SOCIAL BOTS**

The massive spread of fake news has been identified as a major global risk and has been alleged to influence elections and threaten democracies. Communication, cognitive, social, and computer scientists are engaged in efforts to study the complex causes for the viral diffusion of digital misinformation and to develop solutions, while search and social media platforms are beginning to deploy countermeasures. However, to date, these efforts have been mainly informed by anecdotal evidence rather than systematic data.

We find evidence that social bots play a key role in the spread of fake news. Accounts that actively spread misinformation are significantly more likely to be bots. Automated accounts are particularly active in the early spreading phases of viral claims, and tend to target influential users. Humans are vulnerable to this manipulation, retweeting bots who post false news. Successful sources of false and biased claims are heavily supported by social bots. These results suggests that curbing social bots may be an effective strategy for mitigating the spread of online misinformation

**2.4** **MISLEADING ONLINE CONTENT**

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 highlevel, complex abstractions as data representations through a hierarchical learning 7 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

**METHODOLOGY**

**3.1 THE DATASET**

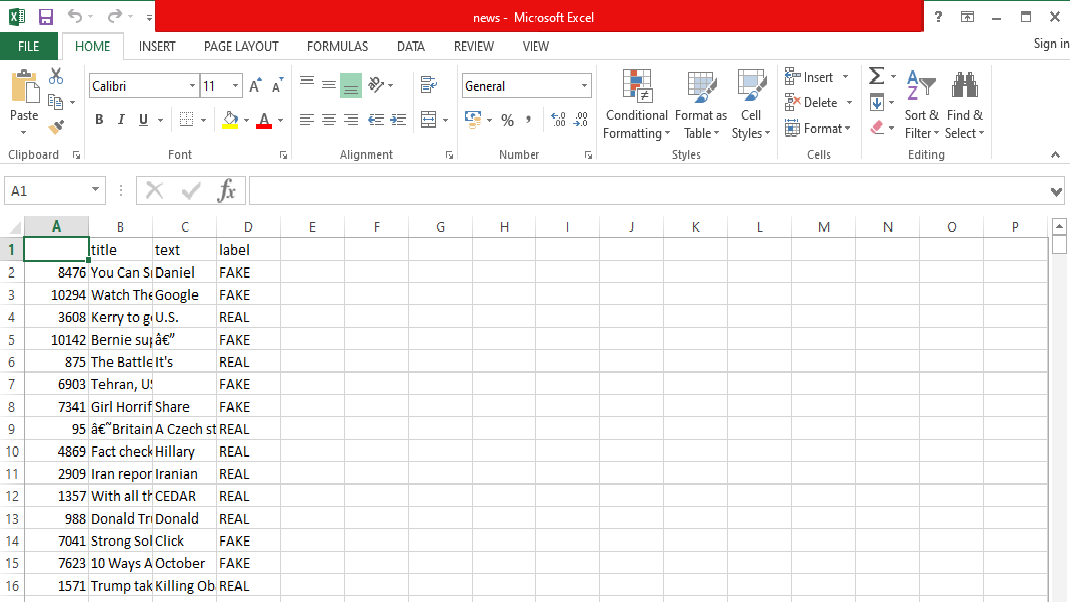
****

Figure 3

The dataset is simple. It contains the titles of the news, the body text and a label field, which, if the news is authentic, shows REAL and if inauthentic, shows FAKE.

**3.2 PROGRAMMING LANGUAGE AND MACHINE LEARNING MODEL**

In this project, Python version 3.5 has been implemented. Python programming language is an open-source programming language and since it is free, its use is extensive and has an active community development and support.

Python programming language offers creation of solutions to machine learning problems with code that is readable and intuitive, its simplicity also enables developers to develop robust, reliable projects. Python is also platform independent which enables the developers to deploy and utilize the code or frameworks on different systems with little to no changes. Python is also

supported by a variety of platforms, some of which includes Windows, MacOS and Linux.

One of the major reasons for implementing Python programming language is its extensive collection of libraries and frameworks. In this project, Pandas, NumPy, Seaborn are a handful of examples of libraries that have enabled developers to create the system quickly and effectively.

There are two parts to the ML Model building. Machine Learning is a part of our life that can help us in predicting. We are using two types of model in this case. For the first part, we used passive-aggressive classifiers. And the steps include:

* **Data Loading**: We are loading a CSV file for the data sorting and training-testing part of the model. The CSV file is turned into an array for easier work purpose.
* **Classifier**: Passive-aggressive algorithms are a family of great learning algorithms. They are similar to Perceptron because it does not require a reading scale. However, unlike Perceptron, they include parameter correction. Passive is used when the prediction is correct and there is no Fake News Detector 16 change in the model. But if there is any kind of change in the model, that is if the prediction is not correct then the aggressive part is called, which changes the model accordingly. The aggressive part of the model changes the model according to its wish on the backend.
* **Model Building:** The model is built through the train and test of the dataset, by ensuring that the training is done for 80% of the dataset and testing is done in the rest of the 20% of the dataset.

In the second part, we used is LSTM. Here are the steps:

1. **Loading the data:** For this step, it is the same as the passive-aggressive one.
2. **Scanning and parsing:** Data is loaded from a CSV file. This consists of the body of selected news articles. It then contains a label field that indicates whether the news is real or fake. In this code block, we scan the CSV and clean the titles to filter out stop words and punctuation.
3. **Tokenization**: The tokenizer is used to assign indices to words, and filter out infrequent words. This allows us to generate sequences for our training and testing data.
4. **Embedding matrix**: Apply the embedding matrix. An embedding matrix is used to extract the semantic information from the words in each title.
5. **Model Building**: Building the model and finding out the accuracy via confusion matrix. The model is created using an Embedding layer, LSTM, Dropout, and Dense layers. We are going to run the data on 20 epochs. We observed that the LSTM model is vastly inaccurate in predicting the authenticity of the news. So we decided to show the output by running it through the Passive-aggressive classifier model.

**3.3 COST BENEFIT ANALYSIS**

The cost-benefit analysis of the project could be:

* Basic cost of setting up a device such as laptop or mobile device.
* Almost no user cost
* Cost for service provider includes the cost to fulfill software requirements as well as the cost of training large amount of data for real news generation from genuine sites.
* Users can insert news article and analyze the authenticity of news which will help mitigate the effects of fake news.

# CHAPTER – 4

## RESULTS AND DISCUSSION

## KEY INSIGHTS

The passive aggressive model produces 93% accuracy. When we input the news text on the interface, it correctly identifies the news most of the time. We tested this by using news from The Onion. The Onion is a satire ‘news’ portal that posts fake funny news. When we pasted some of the news from the site on our web interface, those were correctly identified as fake. But when we wanted to test the news from BBC or New York Times, those were correctly identified as real. But the accuracy of the LSTM model was much lower, so we went with the Passive Aggressive model to produce output on the interface.

* Algorithm’s accuracy depends on the type and size of your dataset. More the data, more chances of getting correct accuracy.
* Machine learning depends on the variations and relations
* Understanding what is predictable is as important as trying to predict it.
* While making algorithm choice, speed should be a consideration factor.

**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.

**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.

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.

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.

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.

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

**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.

# 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.

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.

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.

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

# CHAPTER – 5 CONCLUSION AND FUTURE ENHANCEMENT

## CONCLUSION AND FUTURE WORK

Many people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has negative impacts on individual people and society. In this paper, an innovative model for fake news detection using machine learning algorithms has been presented. This model takes news events as an input and based on twitter reviews and classification algorithms it predicts the percentage of news being fake or real.

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

# CHAPTER – 6

# APPENDIX

**SOURCE CODE:**

import numpy as np

import pandas as pd

import itertools

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.linear\_model import PassiveAggressiveClassifier

from sklearn.metrics import accuracy\_score, confusion\_matrix

df=pd.read\_csv('D:\\DataFlair\\news.csv')

df.shape

df.head()

labels=df.label

labels.head()

x\_train,x\_test,y\_train,y\_test=train\_test\_split(df['text'], labels, test\_size=0.2, random\_state=7)

tfidf\_vectorizer=TfidfVectorizer(stop\_words='english', max\_df=0.7)

tfidf\_train=tfidf\_vectorizer.fit\_transform(x\_train)

tfidf\_test=tfidf\_vectorizer.transform(x\_test)

pac=PassiveAggressiveClassifier(max\_iter=50)

pac.fit(tfidf\_train,y\_train)

y\_pred=pac.predict(tfidf\_test)

score=accuracy\_score(y\_test,y\_pred)

print(f'Accuracy: {round(score\*100,2)}%')

confusion\_matrix(y\_test,y\_pred, labels=['FAKE','REAL'])