

A PROJECT REPORT ON
A SUPERVISED LEARNING ESTIMATOR AND NLP
FRAMEWORK FOR DETECTION OF CYBER FRAUD IN
PROMOTIONAL CONTENT

Submitted in partial fulfilment of the requirements for the award of
the degree of

BACHELOR OF TECHNOLOGY IN
COMPUTER SCIENCE AND ENGINEERING
BY

- | | |
|---------------------------------|-------------------|
| 1. SOURAPU LAL HARINI | 21F21A0527 |
| 2. RAJULA JEEVAMANI | 21F21A0535 |
| 3. ADIMULAM KEERTHANA | 22F25A0505 |
| 4. EDIGA JHANSI | 21F21A0537 |
| 5. VADLA BHAVANI SHANKAR | 21F21A0514 |

Under the Esteemed Guidance of



Dr.P.NAMRATHA M.Tech., PhD.,

Professor & HOD

Department of Computer science And Engineering,

GATES Institute of Technology,

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

	<p style="text-align: center;">GATES INSTITUTE OF TECHNOLOGY (Approved By A.I.C.T.E. New Delhi, Affiliated To J.N.T.U) NH – 7, GOOTY – 515401, ANANTAPUR (D), A.P. www.gatesit.ac.in contact :08552- 252444 2021-2025</p>	
---	--	---

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GATES INSTITUTE OF TECHNOLOGY, GOOTY

CERTIFICATE

This is to certify that the Project Report Entitled "A SUPERVISED LEARNING ESTIMATOR AND NLP FRAMEWORK FOR DETECTION OF CYBER FRAUD IN PROMOTIONAL CONTENT" That is Being Submitted by SOURAPU LAL HARINI (21F21A0527), in the partial fulfillment of requirements for the award of Degree of Bachelor of Technology in **Computer Science and Engineering** From The Gates Institute of Technology is a record of bonafide work carried out by them under my guidance and supervision.

Project Guide:

Dr.P.Namratha, M. Tech.,ph.D.,
Professor & Head,
Department of CSE,
GATES Institute of Technology,
Gooty-515401

Head of the Department:

Dr.P. Namratha, M. Tech, ph. D,
Professor & Head,
Department of CSE,
GATES Institute Of Technology,
Gooty-515401

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We are very much thankful to our beloved correspondent **Smt V.K. Padmavathamma** Garu for providing the necessities for completion of the course.

We wish to thank **Sri. G. Raghunatha Reddy** sir our beloved Managing Director & **Smt. V.K. Srivani Director** providing us with all the facilities that are required for our project.

We cordially thank our Principal **Dr. A. Sudhakar** M.Tech., Ph.D., for providing the necessities in completion of this project.

We wish to thank our HOD **Dr .P. Namratha** M.Tech., Ph.D., for providing us all the facilities that are required for completing of our project.

We express our sincere thanks to our project guide **Dr .P.NAMRATHA**, M.Tech., Ph.D., Of Computer Science and Engineering department, GATES Institute of Technology, Gooty, Anantapur, without whose innovative and imaginative approach, regular monitoring, and timely advice, this project would never have been realized.

We thank our teaching and non-teaching staff of the Department of Computer Science & Engineering, GATES Institute of Technology, Gooty, Anantapur.

Project Associates:

- | | |
|--------------------------|------------|
| 1. SOURAPU LAL HARINI | 21F21A0527 |
| 2. RAJULA JEEVAMANI | 21F21A0535 |
| 3. ADIMULAM KEERTHANA | 22F25A0505 |
| 4. EDIGA JHANSI | 21F21A0537 |
| 5. VADLA BHAVANI SHANKAR | 21F21A0514 |

TABLE OF CONTENTS

ABSTRACT	I
LIST OF FIGURES	II
LIST OF ABBREVIATIONS	III
TITLE	PAGE-NO
<hr/>	
1. INTRODUCTION	1-3
1.1 Motivation	2
1.2 Problem Definition	2
1.3 Objective of the Project	3
2. TECHNOLOGIES LEARNT	4-30
3. LITERATURE SURVEY	31-33
3.1 Introduction	32
3.2 Existing System	32
3.3 Disadvantages of Existing System	32
3.4 Proposed System	33
3.4.1 Goals of New System	33
3.5 Summary	33
4. ANALYSIS	34-38
4.1 Introduction	35
4.2 Software Requirement Specification	35
4.2.1 User Requirements	35
4.2.2 Software Requirements	36
4.2.3 Hardware Requirements	37
4.3 Feasibility	37
4.4 Summary	38
5. DESIGN	39-50
5.1 Introduction	40
5.2 Module Description	42-43
5.3 Module Design	43-50
5.4 Summary	50

6. IMPLEMENTATION AND RESULTS	51-63
6.1 Introduction	52-53
6.2 Method of implementation	55
6.3 Sample Code	55-58
6.4 Output Screens	59-62
7. TESTING AND VALIDATION	63-67
7.1 Introduction	64
7.2 Types of Tests	64
7.2.1 Unit Testin	64
7.2.2 Integration Testing	64
7.2.3 Functional Testing	65
7.2.4 System Testing	65
7.2.5 White Box Testing	65
7.2.6 Black Box Testing	66
7.2.7 Unit Testing	66-67
7.2.8 Acceptance Testing	67
8. CONCLUSION	68-69
9. REFERENCES	70-74

ABSTRACT

The dissemination of intentionally deceptive content disguised as legitimate journalism is a global issue undermining information accuracy and integrity. This phenomenon significantly influences public opinion, decision-making, and voting patterns. Most false news originates on social media platforms like Facebook and Twitter, later infiltrating mainstream media outlets such as television and radio. These false stories often share linguistic traits, including excessive use of unsubstantiated hyperbole and non-attributed quotes. This paper presents the findings of a study on false news detection, focusing on a novel classifier developed using Textblob, Natural Language Toolkit (NLTK), and SciPy. The system employs quoted attribution as a key feature within a Bayesian machine learning framework to estimate the likelihood of an article being false. The methodology achieved a precision rate of 63.333% in identifying false articles containing quotes. This innovative process, termed "influence mining," is introduced as a potential tool for detecting false news and propaganda. The study details the research process, technical analysis, linguistic features, and classifier performance. It concludes with insights into the evolution of the current system into a comprehensive influence mining framework capable of addressing broader disinformation challenges.

LIST OF FIGURES

FIGURE.NO	FIGURE NAME	PAGE NO
1.	2.1 Official Site of python	23
2.	2.2 Python Download Tab	23
3.	2.3 Versions of Python	24
4.	2.4 Files of python	24
5.	2.5 Open Python	25
6.	2.6 Installation Of python	26
7.	2.7 Set Up of Python	26
8.	2.8 Search For Cmd	27
9.	2.9 Open Cmd	28
10.	2.10 Python IDLE	28
11.	2.11 Python File	28
12.	2.12 Python Program	29
13.	5.2.1 System Architecture	43
14.	5.3.2 Use Case Diagram	45
15.	5.3.3 Sequence Diagram	46
16.	5.3.4 Class Diagram	47
17.	5.3.5 Data Flow Diagram	48
18.	5.3.6 Component Diagram	49
19.	5.3.7 Activity Diagram	50
20.	6.4.1 Home Page	59
21.	6.4.2 User Login Page	59
22.	6.4.3 Upload News Article Page	60
23.	6.4.4 Upload News Document Page	60
24.	6.4.5 Uploading Dataset	61
25.	6.4.6 After Uploading Dataset	61
26.	6.4.7 Showing Either REAL OR FLASE News	62

LIST OF ABBREVIATIONS

S.NO	WORD	ABBREVIATION
1	NLP	Natural Language Processing
2	NER	Named Entity Recognition
3	ML	Machine Learning
4	AI	Artificial Intelligence
5	CSV	Comma-Separated-Values
6	UI	User Interface
7	UML	Unified Modeling Language
8	API	Application Programming Interface
9	CPU	Central Processing Unit
10	GUI	Graphical User Interface
11	POS	Part-of-Speech (Tagging)
12	I/O	Input/Output
13	DB	Database
14	DFD	Data Flow Diagram
15	SOM	Self-Organizing Maps

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

1.1. Motivation:

Intentionally deceptive content presented under the guise of legitimate journalism (or ‘False News,’ as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most false news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The false news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a false news classifier are presented and discussed in this paper.

1.2. Problem Definition:

In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system. The false news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a false news identification study that documents the performance of a false news classifier are presented and discussed in this paper.

1.3. Objective of Project:

False news has been demonstrated to be problematic in multiple ways. It has been shown to have real influence on public perception and the ability to shape regional and national dialogue. It has harmed businesses and individuals and even resulted in death, when an individual responded to a hoax. It has caused some teenagers to reject the concept of media objectivity and many students can’t reliably tell the difference between real and false articles. It is even thought to have influenced the 2016 United States elections.

False news can be spread deliberately by humans or indiscriminately by bot armies, with the latter giving a nefarious article significant reach. Not just articles are false, in many cases false, mislabeled or deceptive images are also used to maximize impact. Some contend that fake news is a “plague” on society’s digital infrastructure. Many are working to combat it. Farajtabar, et al., For example, has proposed a system based on points, while Haigh, Haigh and Kozak have suggested the use of “peer-to-peer” counter propaganda.

CHAPTER – 2

TECHNOLOGIES LEARNT

2. TECHNOLOGIES LEARNT

2.1 INTRODUCTION:

What is Python :-

Below are some facts about Python.

- Python is currently the most widely used multi-purpose, high-level programming language
- Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.
- Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.
- Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber... etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

1. [Machine Learning](#)
2. GUI Applications (like [Kivy](#), Tkinter, PyQt etc.)
3. Web frameworks like [Django](#) (used by YouTube, Instagram, Dropbox)
4. image processing (like [OpenCV](#), Pillow)
5. Web scraping (like Scrapy, BeautifulSoup, Selenium)
6. Test frameworks
7. Multimedia

Advantages of Python :-

Let's see how Python dominates over other languages.

1. Extensive Libraries

Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

2. Extensible

As we have seen earlier, Python can be **extended to other languages**. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add **scripting capabilities** to our code in the other language.

4. Improved Productivity

The language's simplicity and extensive libraries render programmers **more productive** than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

5. IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

6. Simple and Easy

When working with Java, you may have to create a class to print '**Hello World**'. But in Python, just a print statement will do. It is also quite **easy to learn, understand, and code**. This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

7. Readable :

Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and **indentation is mandatory**. This further aids the readability of the code.

8. Object-Oriented :

This language supports both the **procedural and object-oriented** programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the **encapsulation of data** and functions into one.

9. Free and Open-Source :

Like we said earlier, Python is **freely available**. But not only can you **download Python** for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

10. Portable :

When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn't the same with Python. Here, you need to **code only once**, and you can run it anywhere. This is called **Write Once Run Anywhere (WORA)**. However, you need to be careful enough not to include any system-dependent features.

11. Interpreted :

Lastly, we will say that it is an interpreted language. Since statements are executed one by one, **debugging is easier** than in compiled languages.

Advantages of Python Over Other Languages:

1. Less Coding :

Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don't have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.

2. Affordable :

Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.

The 2019 Github annual survey showed us that Python has overtaken Java in the most popular programming language category.

3. Python is for Everyone:

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and **machine learning**, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

Disadvantages of Python:

So far, we've seen why Python is a great choice for your project. But if you choose it, you should be aware of its consequences as well. Let's now see the downsides of choosing Python over another language.

1. Speed Limitations :

We have seen that Python code is executed line by line. But since [Python](#) is interpreted, it often results in **slow execution**. This, however, isn't a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

2. Weak in Mobile Computing and Browsers :

While it serves as an excellent server-side language, Python is much rarely seen on the **client-side**. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called **Carbonnelle**.

The reason it is not so famous despite the existence of Brython is that it isn't that secure.

3. Design Restrictions :

As you know, Python is **dynamically-typed**. This means that you don't need to declare the type of variable while writing the code. It uses **duck-typing**. But wait, what's that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can **raise run-time errors**.

4. Underdeveloped Database Access Layers :

Compared to more widely used technologies like **JDBC (Java DataBase Connectivity)** and **ODBC (Open DataBase Connectivity)**, Python's database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

5. Simple :

No, we're not kidding. Python's simplicity can indeed be a problem. Take my example. I don't do Java, I'm more of a Python person. To me, its syntax is so simple that the verbosity of Java code seems unnecessary.

This was all about the Advantages and Disadvantages of Python Programming Language.

History of Python : -

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde & Informatica). The greatest achievement of ABC was to influence the design of Python. Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners¹, Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde en Informatica (CWI). I don't know how well people know ABC's influence on Python. I try to mention ABC's influence because I'm indebted to everything I learned during that project and to the people who worked on it." Later on in the same Interview, Guido van Rossum continued: "I remembered all my experience and some of my frustration with ABC. I decided to try to design a simple scripting language that possessed some of ABC's better properties, but without its problems. So I started typing. I created a simple virtual machine, a simple parser, and a simple runtime. I made my own version of the various ABC parts that I liked. I created a basic syntax, used indentation for statement grouping instead of curly braces or begin-end blocks, and developed a small number of powerful data types: a hash table (or dictionary, as we call it), a list, strings, and numbers."

What is Machine Learning : -

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of *building models of data*. Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models *tunable parameters* that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be program and learning of

used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain. Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

Categories Of Machine Learning :-

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

- *Supervised learning* involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into *classification* tasks and *regression* tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.
- *Unsupervised learning* involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as *clustering* and *dimensionality reduction*. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

Need for Machine Learning

Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and haven't surpassed human intelligence in many aspects. Then the question is that what

is the need to make machine learn? The most suitable reason for doing this is, “to make decisions, based on data, with efficiency and scale”.Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programing logic, in the problems that cannot be programmed inherently. The fact is that we can’t do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

Challenges in Machines Learning :-

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are –

Quality of data – Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

Time-Consuming task – Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

Lack of specialist persons – As ML technology is still in its infancy stage, availability of expert resources is a tough job.

No clear objective for formulating business problems – Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

Issue of overfitting & underfitting – If the model is overfitting or underfitting, it cannot be represented well for the problem.

Curse of dimensionality – Another challenge ML model faces is too many features of data points. This can be a real hindrance.

Difficulty in deployment – Complexity of the ML model makes it quite difficult to be deployed in real life.

Applications of Machines Learning :-

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML –

- Emotion analysis
- Sentiment analysis
- Error detection and prevention
- Weather forecasting and prediction
- Stock market analysis and forecasting
- Speech synthesis
- Speech recognition
- Customer segmentation
- Object recognition
- Fraud detection
- Fraud prevention
- Recommendation of products to customer in online shopping

How to Start Learning Machine Learning?

Arthur Samuel coined the term “**Machine Learning**” in 1959 and defined it as a “**Field of study that gives computers the capability to learn without being explicitly programmed**”.

And that was the beginning of Machine Learning! In modern times, Machine Learning is one of the most popular (if not the most!) career choices. According to [Indeed](#), Machine Learning

Engineer Is The Best Job of 2019 with a 344% growth and an average base salary of **\$146,085** per year. But there is still a lot of doubt about what exactly is Machine Learning and how to start learning it? So this article deals with the Basics of Machine Learning and also the path you can follow to eventually become a full-fledged Machine Learning Engineer. Now let's get started!!!

How to start learning ML?

This is a rough roadmap you can follow on your way to becoming an insanely talented Machine Learning Engineer. Of course, you can always modify the steps according to your needs to reach your desired end-goal!

Step 1 – Understand the Prerequisites :In case you are a genius, you could start ML directly but normally, there are some prerequisites that you need to know which include Linear Algebra, Multivariate Calculus, Statistics, and Python. And if you don't know these, never fear! You don't need a Ph.D. degree in these topics to get started but you do need a basic understanding

(a) Learn Linear Algebra and Multivariate Calculus

Both Linear Algebra and Multivariate Calculus are important in Machine Learning. However, the extent to which you need them depends on your role as a data scientist. If you are more focused on application heavy machine learning, then you will not be that heavily focused on maths as there are many common libraries available. But if you want to focus on R&D in Machine Learning, then mastery of Linear Algebra and Multivariate Calculus is very important as you will have to implement many ML algorithms from scratch.

(b) Learn Statistics

Data plays a huge role in Machine Learning. In fact, around 80% of your time as an ML expert will be spent collecting and cleaning data. And statistics is a field that handles the collection, analysis, and presentation of data. So it is no surprise that you need to learn it!!! Some of the key concepts in statistics that are important are Statistical Significance, Probability Distributions, Hypothesis Testing, Regression, etc. Also, Bayesian Thinking is also a very important part of ML which deals with various concepts like Conditional Probability, Priors, and Posteriors, Maximum Likelihood, etc.

(c) Learn Python

Some people prefer to skip Linear Algebra, Multivariate Calculus and Statistics and learn them as they go along with trial and error. But the one thing that you absolutely cannot skip is [Python](#)! While there are other languages you can use for Machine Learning like R, Scala, etc. Python is currently the most popular language for ML. In fact, there are many Python libraries that are specifically useful for Artificial Intelligence and Machine Learning such as [Keras](#), [TensorFlow](#), [Scikit-learn](#), etc. So if you want to learn ML, it's best if you learn Python! You can do that using various online resources and courses such as **Fork Python** available Free on GeeksforGeeks.

Step 2 – Learn Various ML Concepts

Now that you are done with the prerequisites, you can move on to actually learning ML (Which is the fun part!!!) It's best to start with the basics and then move on to the more complicated stuff. Some of the basic concepts in ML are:

(a) Terminologies of Machine Learning

- **Model** – A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called a hypothesis.
- **Feature** – A feature is an individual measurable property of the data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, etc.
- **Target (Label)** – A target variable or label is the value to be predicted by our model. For the fruit example discussed in the feature section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.
- **Training** – The idea is to give a set of inputs(features) and it's expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.
- **Prediction** – Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output(label).

(b) Types of Machine Learning

- **Supervised Learning** – This involves learning from a training dataset with labeled data using classification and regression models. This learning process continues until the required level of performance is achieved.
- **Unsupervised Learning** – This involves using unlabelled data and then finding the underlying structure in the data in order to learn more and more about the data itself using factor and cluster analysis models.
- **Semi-supervised Learning** – This involves using unlabelled data like Unsupervised Learning with a small amount of labeled data. Using labeled data vastly increases the learning accuracy and is also more cost-effective than Supervised Learning.
- **Reinforcement Learning** – This involves learning optimal actions through trial and error. So the next action is decided by learning behaviors that are based on the current state and that will maximize the reward in the future.

Advantages of Machine learning :-**1. Easily identifies trends and patterns -**

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. Continuous Improvement

As **ML algorithms** gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

Disadvantages of Machine Learning :-

1. Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2. Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

3. Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4. High error-susceptibility

[Machine Learning](#) is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

Python Development Steps : -

Guido Van Rossum published the first version of Python code (version 0.9.0) at alt.sources in February 1991. This release included already exception handling, functions, and the core data types of list, dict, str and others.

It was also object oriented and had a module system. Python version 1.0 was released in January 1994. The major new features included in this release were the functional programming tools lambda, map, filter and reduce, which Guido Van Rossum never liked. Six and a half years later in October 2000, Python 2.0 was introduced. This release included list comprehensions, a full garbage collector and it was supporting unicode. Python flourished for another 8 years in the versions 2.x before the next major release as Python 3.0 (also known as "Python 3000" and "Py3K") was released. Python 3 is not backwards compatible with Python 2.x. The emphasis in Python 3 had been on the removal of duplicate programming constructs and modules, thus fulfilling or coming close to fulfilling the 13th law of the Zen of Python: "There should be one -- and preferably only one -- obvious way to do it." Some changes in Python 7.3

- Print is now a functions
- Views and iterators instead of lists
- The rules for ordering comparisons have been simplified. E.g. a heterogeneous list cannot be sorted, because all the elements of a list must be comparable to each other.
- There is only one integer type left, i.e. int. long is int as well.
- The division of two integers returns a float instead of an integer. "//" can be used to have the "old" behaviour.

Purpose :-

We demonstrated that our approach enables successful segmentation of intra-retinal layers—even with low-quality images containing speckle noise, low contrast, and different intensity ranges throughout—with the assistance of the ANIS feature.

Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

- Python is Interpreted – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive – you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

Modules Used in Project :-

Tensorflow

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas: Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains

Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

Scikit – learn: Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

- Python is Interpreted – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- Python is Interactive – you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python also acknowledges that speed of development is important. Readable and terse code is part of this, and so is access to powerful constructs that avoid tedious repetition of code. Maintainability also ties into this may be an all but useless metric, but it does say something about

how much code you have to scan, read and/or understand to troubleshoot problems or tweak behaviors. This speed of development, the ease with which a programmer of other languages can pick up basic Python skills and the huge standard library is key to another area where Python excels. All its tools have been quick to implement, saved a lot of time, and several of them have later been patched and updated by people with no Python background - without breaking.

Install Python Step-by-Step in Windows and Mac :

Python a versatile programming language doesn't come pre-installed on your computer devices. Python was first released in the year 1991 and until today it is a very popular high-level programming language. Its style philosophy emphasizes code readability with its notable use of great whitespace.

The object-oriented approach and language construct provided by Python enables programmers to write both clear and logical code for projects. This software does not come pre-packaged with Windows.

How to Install Python on Windows and Mac :

There have been several updates in the Python version over the years. The question is how to install Python? It might be confusing for the beginner who is willing to start learning Python but this tutorial will solve your query. The latest or the newest version of Python is version 3.7.4 or in other words, it is Python 3.

Note: The python version 3.7.4 cannot be used on Windows XP or earlier devices.

Before you start with the installation process of Python. First, you need to know about your **System Requirements**. Based on your system type i.e. operating system and based processor, you must download the python version. My system type is a **Windows 64-bit operating system**. So the steps below are to install python version 3.7.4 on Windows 7 device or to install Python 3. [Download the Python Cheatsheet here.](#) The steps on how to install Python on Windows 10, 8 and 7 are **divided into 4 parts** to help understand better.

Download the Correct version into the system

Step 1: Go to the official site to download and install python using Google Chrome or any other web browser. OR Click on the following link: <https://www.python.org>



Figure 2.1 Official site of python

Now, check for the latest and the correct version for your operating system.

Step 2: Click on the Download Tab.



Figure 2.2 Python Download Tab

Step 3: You can either select the Download Python for windows 3.7.4 button in Yellow Color or you can scroll further down and click on download with respective to their version. Here, we are downloading the most recent python version for windows 3.7.4

Looking for a specific release?

Python releases by version number:

Release version	Release date	Click for more	
Python 3.7.4	July 8, 2019	Download	Release Notes
Python 3.6.9	July 2, 2019	Download	Release Notes
Python 3.7.3	March 25, 2019	Download	Release Notes
Python 3.4.10	March 19, 2019	Download	Release Notes
Python 3.5.7	March 19, 2019	Download	Release Notes
Python 2.7.16	March 4, 2019	Download	Release Notes
Python 3.7.2	Dec. 24, 2018	Download	Release Notes

Figure 2.3 Versions Of Python

Step 4: Scroll down the page until you find the Files option.

Step 5: Here you see a different version of python along with the operating system.

Files					
Version	Operating System	Description	MD5 Sum	File Size	GPU
Stipped source tarball	Source release		68111671e5b2db4ae77b9ab01b7079be	23017643	3xG
XZ compressed source tarball	Source release		d33e4aae66097051c3eca45ee3604803	17133432	3xG
macOS 64-bit/32-bit installer	Mac OS X	for Mac OS X 10.5 and later	6428b4fa7583da71a442c8a8ce08e6	34898416	3xG
macOS 64-bit installer	Mac OS X	for OS X 10.9 and later	5dd605c38217a45773bf5e4a936b2a3f	28882845	3xG
Windows .hug file	Windows		063999573a2c982ac58ade0b4f7cd2	8131761	3xG
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64	9b063bf5d5ee0b0a8e02184a03728a2	7504391	3xG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64	a702b4b0a070de5db3043a583e5a3400	26883348	3xG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64	28c31c608b6d73ae653a3b0351b4bd2	1362904	3xG
Windows x86 embeddable zip file	Windows		9fab08d18841879fda94132574139d8	674628	3xG
Windows x86 executable installer	Windows		33c0802942a5444a3d8451478394788	25663848	3xG
Windows x86 web-based installer	Windows		1b670c1e5d3117d82c30983ea371d87c	1324608	3xG

Figure 2.4 Files Of Python

- To download **Windows 32-bit python**, you can select any one from the three options: Windows x86 embeddable zip file, Windows x86 executable installer or Windows x86 web-based installer.
- To download **Windows 64-bit python**, you can select any one from the three options: Windows x86-64 embeddable zip file, Windows x86-64 executable installer or Windows x86-64 web-based installer.

Here we will install Windows x86-64 web-based installer. Here your first part regarding which version of python is to be downloaded is completed. Now we move ahead with the second part in installing python i.e. Installation

Note: To know the changes or updates that are made in the version you can click on the Release Note Option.

Installation of Python

Step 1: Go to Download and Open the downloaded python version to carry out the installation process.

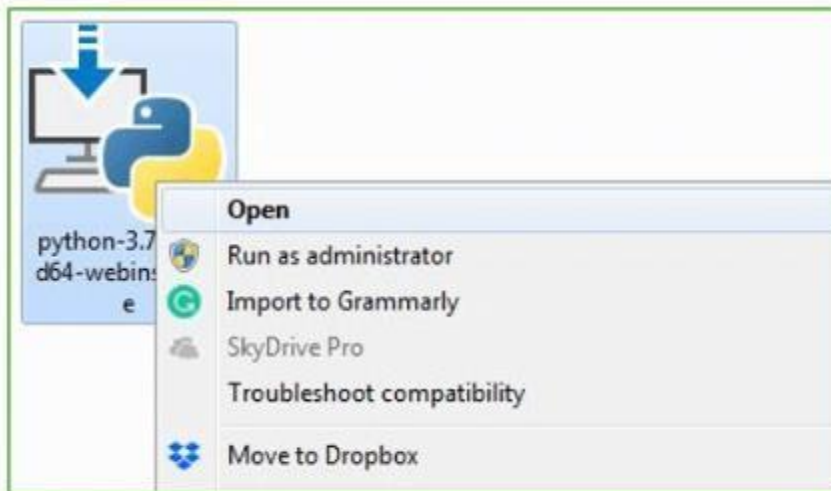


Figure 2.5 Open Python

Step 2: Before you click on Install Now, Make sure to put a tick on Add Python 3.7 to PATH.



Figure 2.6 Installion Of Python

Step 3: Click on Install NOW After the installation is successful. Click on Close.



Figure 2.7 Set Up Python

With these above three steps on python installation, you have successfully and correctly installed Python. Now is the time to verify the installation.

Note: The installation process might take a couple of minutes.

Verify the Python Installation

Step 1: Click on Start

Step 2: In the Windows Run Command, type “cmd”

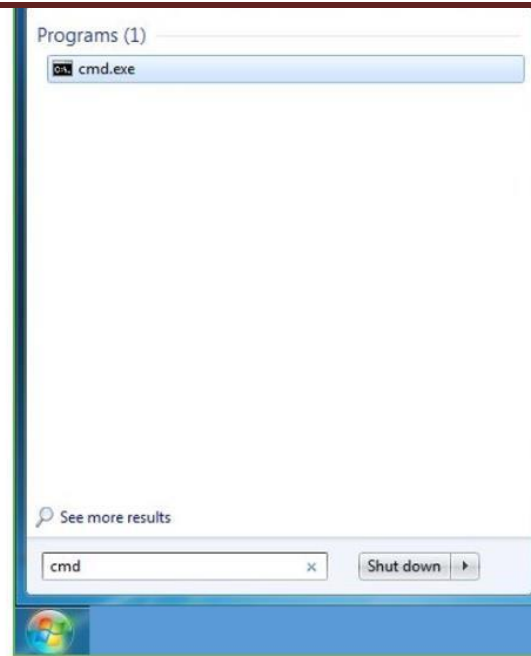


Figure 2.8 Search For Cmd

Step 3: Open the Command prompt option.

Step 4: Let us test whether the python is correctly installed. Type **python -V** and press Enter.

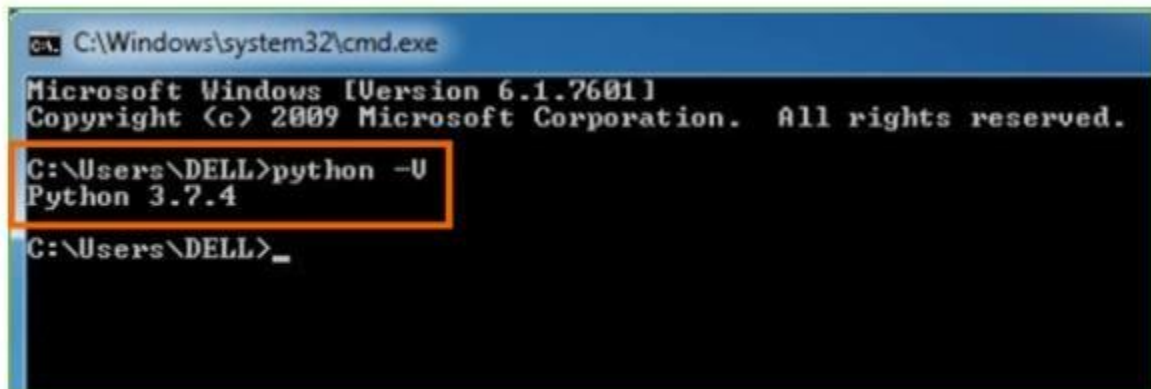


Figure 2.9 Open Cmd

Step 5: You will get the answer as 3.7.4

Note: If you have any of the earlier versions of Python already installed. You must first uninstall the earlier version and then install the new one.

Check how the Python IDLE works

Step 1: Click on Start

Step 2: In the Windows Run command, type “python idle”

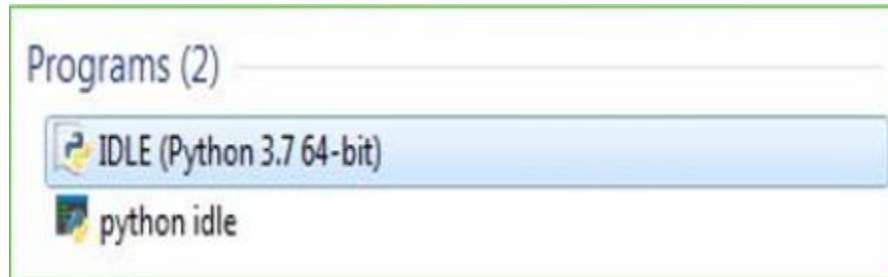


Figure 2.10 Python IDLE

Step 3: Click on IDLE (Python 3.7 64-bit) and launch the program

Step 4: To go ahead with working in IDLE you must first save the file. **Click on File > Click on Save**

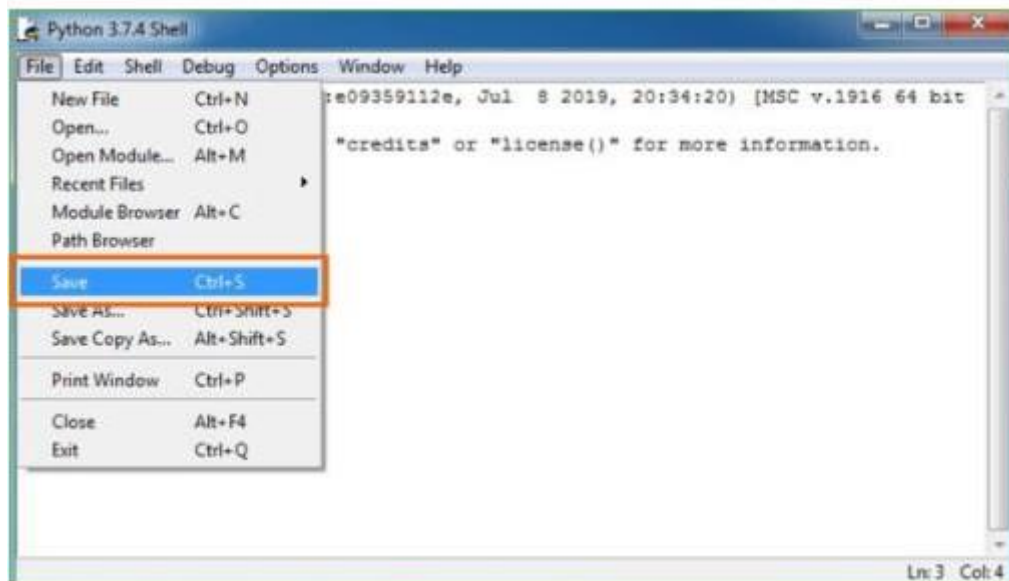


Figure 2.11 Python File

Step 5: Name the file and save as type should be Python files. Click on SAVE. Here I have named the files as Hey World.

Step 6: Now for e.g. **enter print (“Hey World”)** and Press Enter.

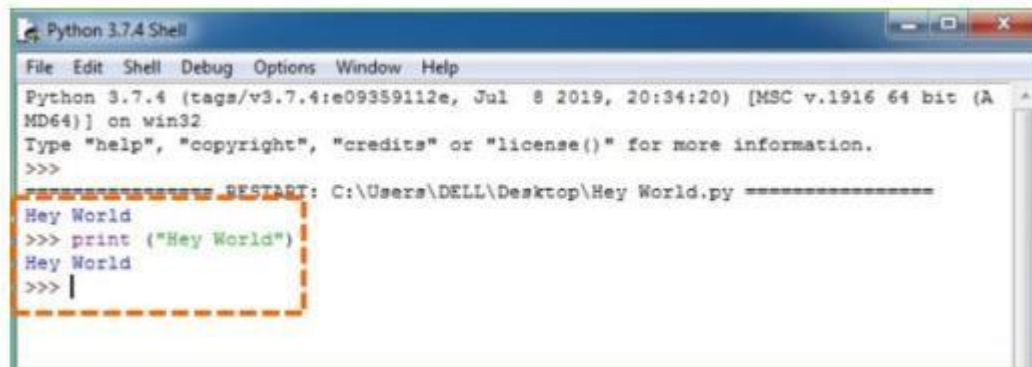
A screenshot of a Python 3.7.4 Shell window. The window has a menu bar with 'File', 'Edit', 'Shell', 'Debug', 'Options', 'Window', and 'Help'. The main text area shows the following content: 'Python 3.7.4 (tags/v3.7.4:09359112e, Jul 8 2019, 20:34:20) [MSC v.1916 64 bit (AMD64)] on win32', 'Type "help", "copyright", "credits" or "license()" for more information.', and a prompt '>>>'. Below the prompt, the text 'Hey World' is displayed. Then, the command '>>> print ("Hey World")' is entered, followed by another prompt '>>>'. The output 'Hey World' is shown again. A dashed orange box highlights the input and output lines. At the top of the window, a status bar reads 'RESTART: C:\Users\DELL\Desktop\Hey World.py'.

Figure 2.12 Python Program

You will see that the command given is launched. With this, we end our tutorial on how to install Python. You have learned how to download python for windows into your respective operating system

CHAPTER-3

LITERATURE SURVEY

3.LITERATURE SURVEY

3.1 Introduction:

The rapid proliferation of misinformation and false news has become a significant societal challenge, especially in the digital age where news dissemination is instantaneous and largely unfiltered. With the increasing reliance on online platforms for information, distinguishing credible content from deceptive narratives has become critical. This issue has drawn considerable attention from researchers in fields such as computer science, journalism, psychology, and linguistics. In recent years, Natural Language Processing (NLP) has emerged as a powerful tool in the fight against fake news. By leveraging machine learning, deep learning, and linguistic analysis, NLP techniques can process vast amounts of textual data to detect patterns indicative of false information. This literature survey explores the various methodologies and models proposed in previous studies for detecting and classifying false news articles. It aims to highlight key contributions, compare different approaches, and identify gaps that this research seeks to address.

3.2 Existing System:

Up until recently, the majority of PDS research has concentrated on securing data stored in the PDS and enforcing user privacy settings. However, the crucial problem of assisting users in defining their privacy settings with PDS data has not yet been thoroughly examined. Because average PDS users lack the necessary skills to comprehend how to convert their privacy requirements into a set of privacy preferences, this is a major problem. Average users may find it challenging to appropriately specify potentially complex privacy choices, as multiple studies have demonstrated.

3.3 Disadvantages of Existing System:

Our digitally generated personal data is dispersed throughout several internet systems run by various providers (banking, airlines, doctors, online social media, etc.). Because each source maintains a different picture of the data, users are unable to fully utilize it while also losing control over their data, which is the responsibility of the data provider.

3.4 Proposed System:

Personal Data Storage (PDS) has inaugurated a substantial change to the way people can store and control their personal data, by moving from a service-centric to a user-centric model. PDSs enable individuals to collect into a single logical vault personal information they are producing. Such data can then be connected and exploited by proper analytical tools, as well as shared with third parties under the control of end users.

3.4.1 Goals of New System:

Advantages Of Proposed System:

- It is desirable to use COX data for phylogenetic exploration.
- We use the data of COX experimental values.
- Security

3.5 Summary:

This project tackles the challenge of fake news by using Natural Language Processing (NLP) and Machine Learning (ML) to classify news articles as real or false. It focuses on in-article attribution features like quotes and named entities, applying supervised learning models such as SVM, Logistic Regression, and Random Forest. The results show that attribution-based features improve classification accuracy, offering a reliable approach to false news detection.

CHAPTER-4

ANALYSIS

4.ANALYSIS

4.1 Introduction:

In the development of any data-driven application. It involves understanding the problem, gathering requirements, and analyzing the functionality of the system to ensure it meets the intended goals. In this project, system analysis is focused on designing an effective false news detection model using Natural Language Processing (NLP) and Machine Learning (ML). The primary goal is to build a system that can accurately classify news articles based on in-article attribution and linguistic patterns

4.2 Software Requirement Specification:

4.2.1 User Requirements:

The system should be designed to provide a seamless user experience while effectively classifying fake news articles. The requirements are divided into functional, non-functional, software, and hardware categories.

1. Functional Requirements :

- The user should be able to upload news articles (in text or CSV format).
- The system should preprocess the uploaded content (tokenization, stop-word removal, stemming).
- It must extract features like named entities, verbs, and quotes using NLP.
- The system should apply a supervised learning model to classify the news as real or false.
- Users should view the classification results with a confidence score.

2. Non-Functional Requirements

- The system should be responsive and provide results in real-time or with minimal delay.
- It should be easy to use, even for non-technical users.
- The application must ensure user data privacy and prevent unauthorized access.
- The system should be maintainable, scalable, and adaptable for future improvements.

4.2.2 Software Requirements:

Functional requirements for a secure cloud storage service are straightforward:

1. The service should be able to store the user's data;
2. The data should be accessible through any devices connected to the Internet;
3. The service should be capable to synchronize the user's data between multiple devices (notebooks, smart phones, etc.);
4. The service should preserve all historical changes (versioning);
5. Data should be shareable with other users;
6. The service should support SSO; and
7. The service should be interoperable with other cloud storage services, enabling data migration from one CSP to another.

• **Operating System:** Windows

• **Coding Language:** Python 3.7

• **Libraries/Frameworks:**

1. TensorFlow
2. Scikit-learn

3. Pandas
4. NumPy
5. NLTK / SpaCy

- **Database (if needed):** SQLite / MySQL (optional for storing results or logs)

4.2.3 Hardware Requirements:

The selection of Hardware Is very important in the existence and performance of any software. The size and capacity are the main requirements. The typical web server must have the following specifications for good performance:

- **Processor** - Pentium –III
- **Speed** – 2.4 GHz
- **RAM** - 512 MB (min)
- **Hard Disk** - 20 GB
- **Floppy Drive** - 1.44 MB
- **Key Board** - Standard Keyboard
- **Monitor** – 15 VGA Colour

4.3 Feasibility:

A feasibility study evaluates the practicality and viability of a proposed system before its development. For the false news classification system using NLP and ML, the following feasibility aspects are considered:

Three key considerations involved in the feasibility analysis are

- ❖ Economical Feasibility
- ❖ Technical Feasibility
- ❖ Social Feasibility
- ❖ Legal and Ethical Feasibility

Economical feasibility:

The use of open-source software significantly reduces development and operational costs. No major financial investment is needed beyond basic computing infrastructure. Once developed, maintenance costs are minimal, and updates can be handled with ease.

Technical feasibility:

The project leverages well-established technologies such as Python, Scikit-learn, and NLP libraries like NLTK and SpaCy. It does not require specialized hardware; a standard computer with moderate specifications is sufficient. Pre-trained models (e.g., BERT, TF-IDF) can be integrated easily, reducing the need for training large datasets from scratch. Developers can implement, test, and deploy the system using freely available open-source tools.

Operational feasibility:

The system is user-friendly, requiring minimal interaction from users—primarily uploading or inputting text. It addresses a critical real-world problem, ensuring user acceptance and usefulness. It can be used by individuals, journalists, or institutions to verify article credibility.

Legal and Ethical Feasibility:

The system does not infringe on any user data privacy rights, as it analyzes public or user-provided text data only. Proper attribution and transparency will be ensured in AI decisions to promote ethical AI use.

4.4 Summary:

Understanding the properties and requirements of a new system is more difficult and requires creative thinking. Understanding of existing running system is also difficult, improper understanding of present system can lead to diversion from solution.

CHAPTER-5

DESIGN

5. DESIGN

5.1 Introduction

The design phase is a crucial part of any software project, as it lays the foundation for how the system will function and interact. In this project, the Fake News Detection System is designed to efficiently process news articles, extract important linguistic features using Natural Language Processing (NLP), and classify the content using a supervised learning algorithm.

The goal of the design is to create a system that is modular, easy to use, and accurate in identifying fake news. This involves defining the architecture, data flow, and interactions between different components of the system.

Through the use of UML diagrams such as use case diagrams, sequence diagrams, class diagrams, and activity diagrams, the structure and behavior of the system are clearly illustrated. These design tools help ensure that each module of the system works together smoothly and supports future improvements or scaling

5.2 Module Description

1. **Tensorflow:** TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.
2. **Numpy:** Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:
 - A powerful N-dimensional array object
 - Sophisticated (broadcasting) functions

- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

3. **Pandas:** Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.
4. **Matplotlib:** Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

5. **Scikit-learn:** Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

- **NumPy:** Base n-dimensional array package
- **SciPy:** Fundamental library for scientific computing
- **Matplotlib:** Comprehensive 2D/3D plotting
- **IPython:** Enhanced interactive console
- **Sympy:** Symbolic mathematics
- **Pandas:** Data structures and analysis
- Extensions or modules for SciPy are conventionally named SciKits. As such, the module

Algorithms used in this project :-

CLUSTERING ALGORITHM :

- Clustering is an unendorsed culture algorithm that finds the concealed arrangement in the unlabeled data. In this work, we used the filter the values, adjust the data values, then apply the hierarchical method, k-means algorithm, self-organizing maps (SOM), and finally apply the Principal Component Analysis (PCA) for avoid the unwanted values, adjust the data with help of log transform, for clustering genes and arrays with hierarchical clustering by centroid linkage,
- Clustering is an unsupervised machine learning technique used to group similar data points together without using labels. Think of it like sorting news articles into "buckets" based on their content without knowing which are false or real in advance.

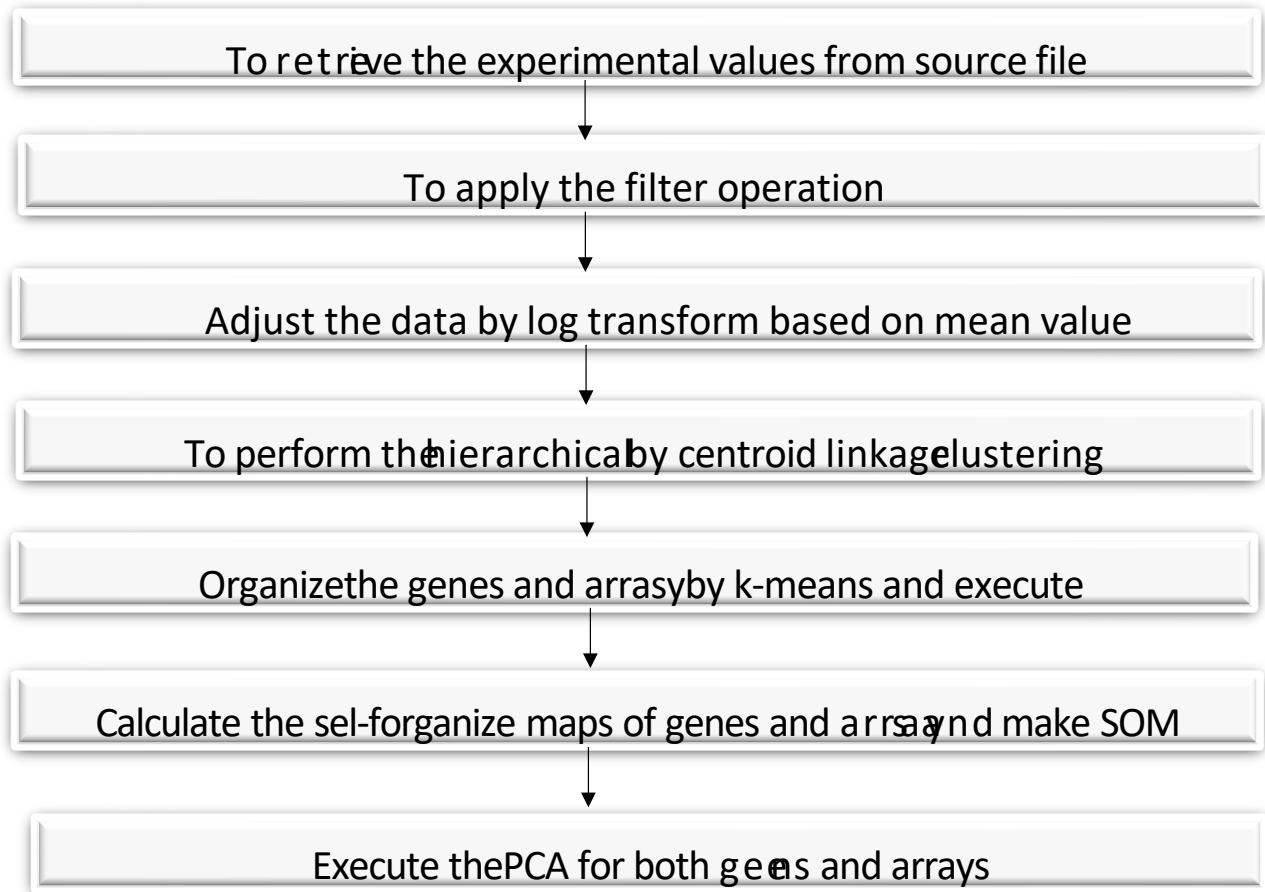


Figure 5.2.1 System Architecture

5.3 Module Design:

- UML is an acronym that stands for **Unified Modeling Language**. Simply put, UML is a modern approach to modeling and documenting software. In fact, it's one of the most popular business process modeling techniques.
- It is based on **diagrammatic representations** of software components. As the old proverb says: "a picture is worth a thousand words". By using visual representations, we are able to better understand possible flaws or errors in software or business processes.
- UML was created as a result of the chaos revolving around software development and documentation. In the 1990s, there were several different ways to represent and document software systems. The need arose for a more unified way to visually represent those systems and as a result, in 1994-1996, the UML was developed by three software engineers working at [Rational Software](#).

- It was later adopted as the standard in 1997 and has remained the standard ever since, receiving only a few updates.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

i. USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

- User – Interacts with the system (uploads file, runs detection)
- System – The backend system that processes the data and classifies news

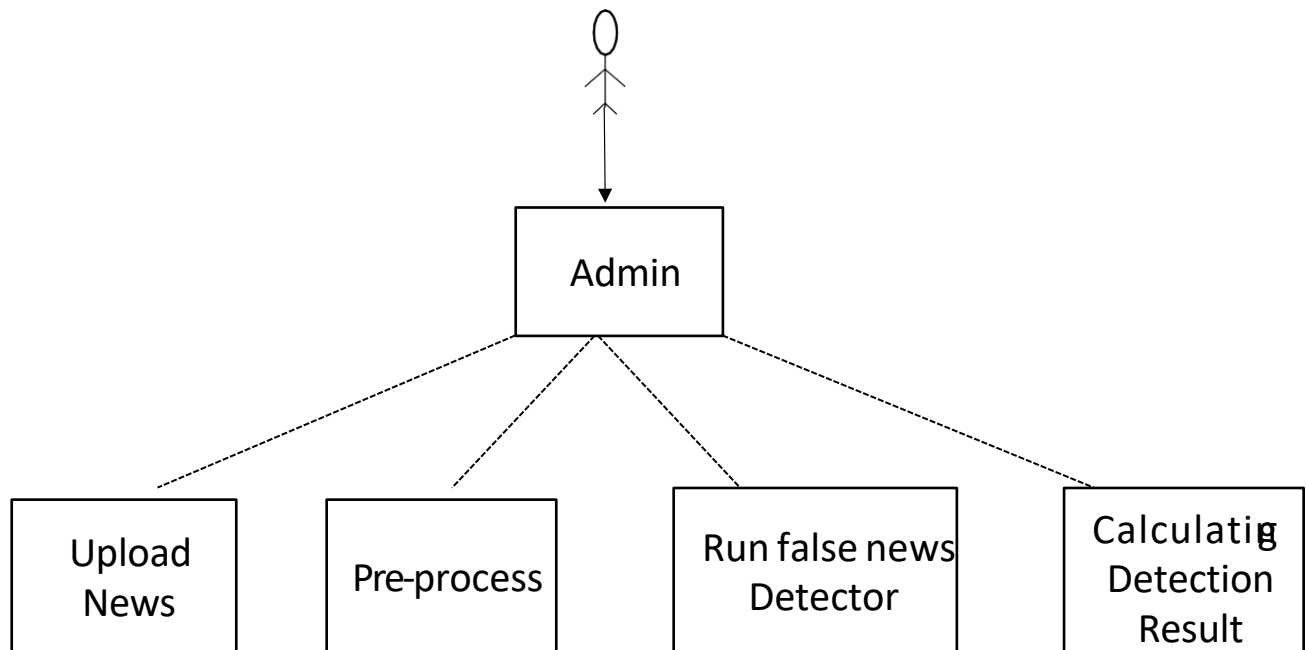


Figure 5.3.2 Activity Diagram

ii. SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

- Actors: Who interacts with the system (e.g., User)
- Objects: Components in the system (Web Interface, Detection System, Database)
- Messages: What actions happen and in what order (like “upload file,” “run classifier”)

The sequence diagram helps visualize how a user action (like uploading a file) flows through the system step by step until the fake/real result is shown.

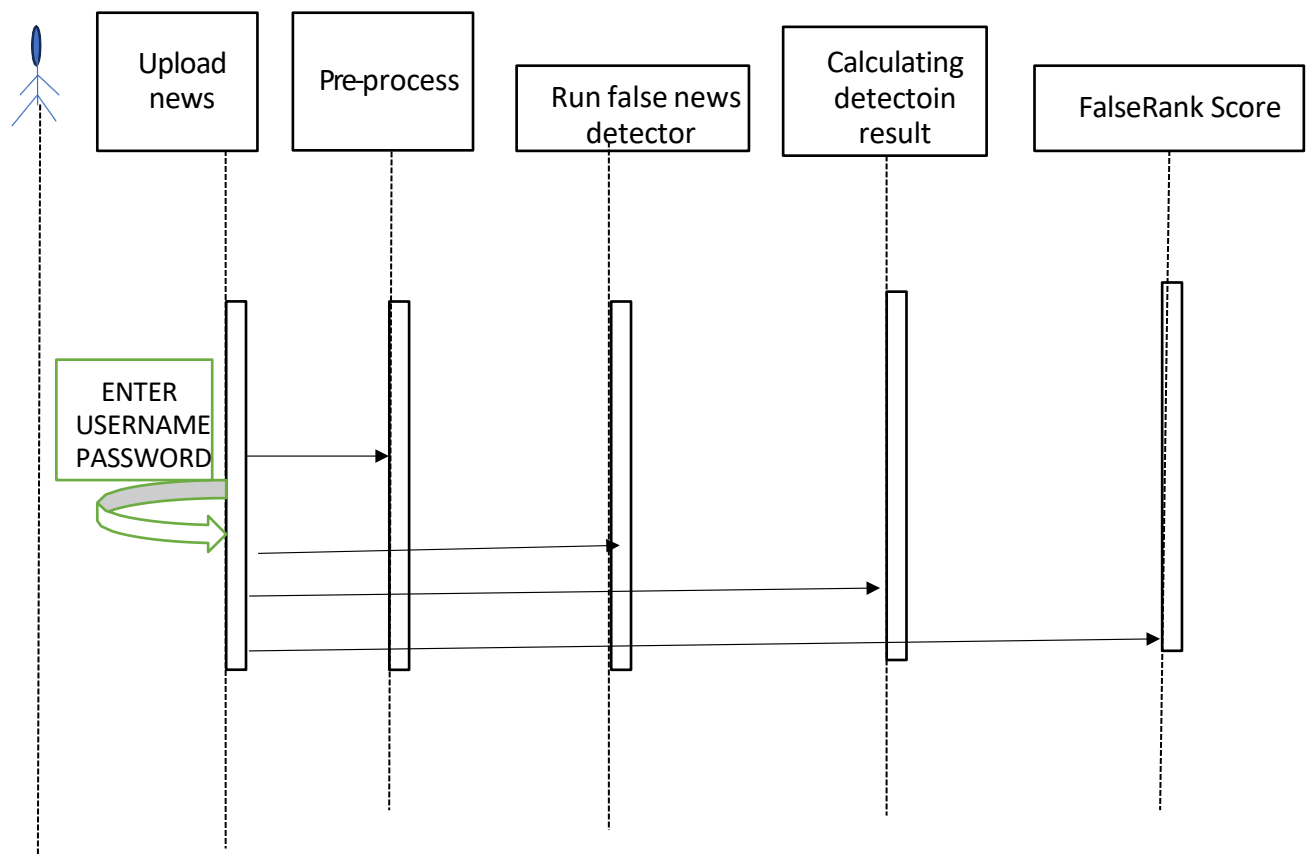


Figure 5.3.3 Sequence Diagram

iii. CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

- Classes (like objects or components in your code)
- Their attributes (data or variables)
- Their methods (functions or actions they can perform)
- The relationships between classes (like inheritance or dependency)

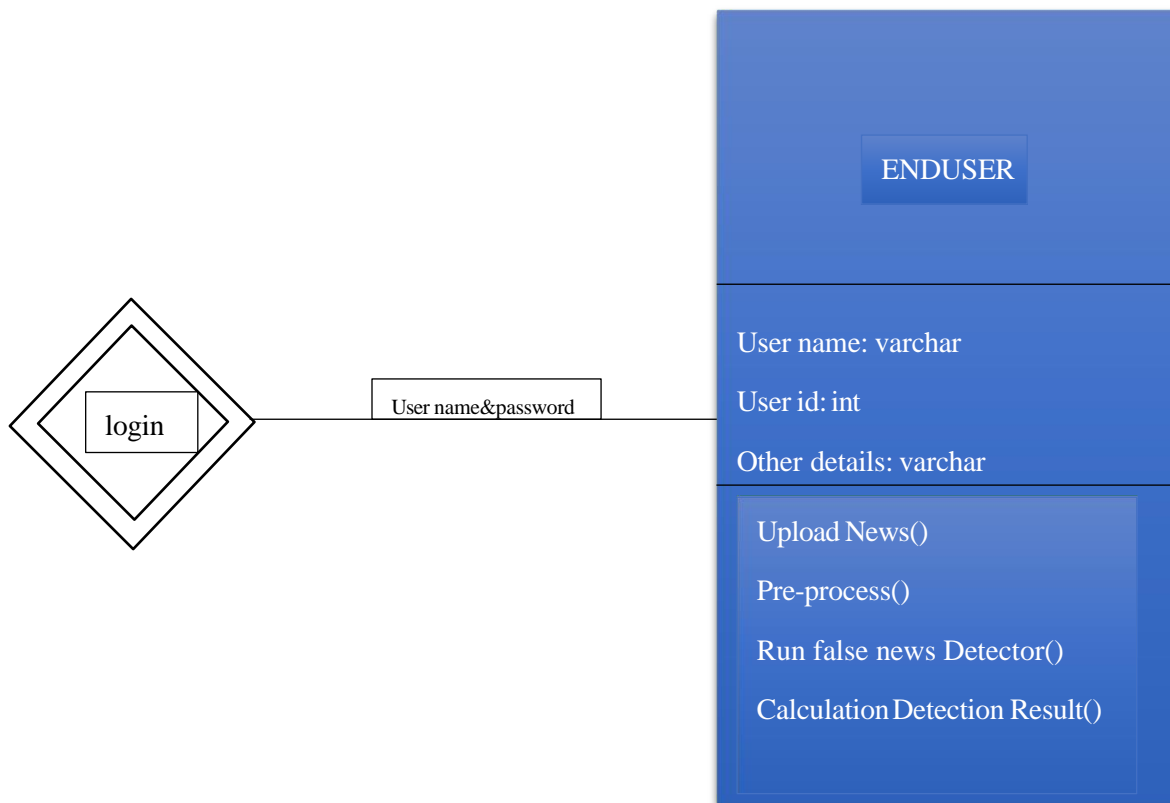


Figure 5.3.4 Class diagram

iv. DATA FLOW DIAGRAM :-

Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.. DFD graphically representing the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. Structure of DFD allows starting from a broad overview and expand it to a hierarchy of detailed diagrams. DFD has often been used due to the following reasons:

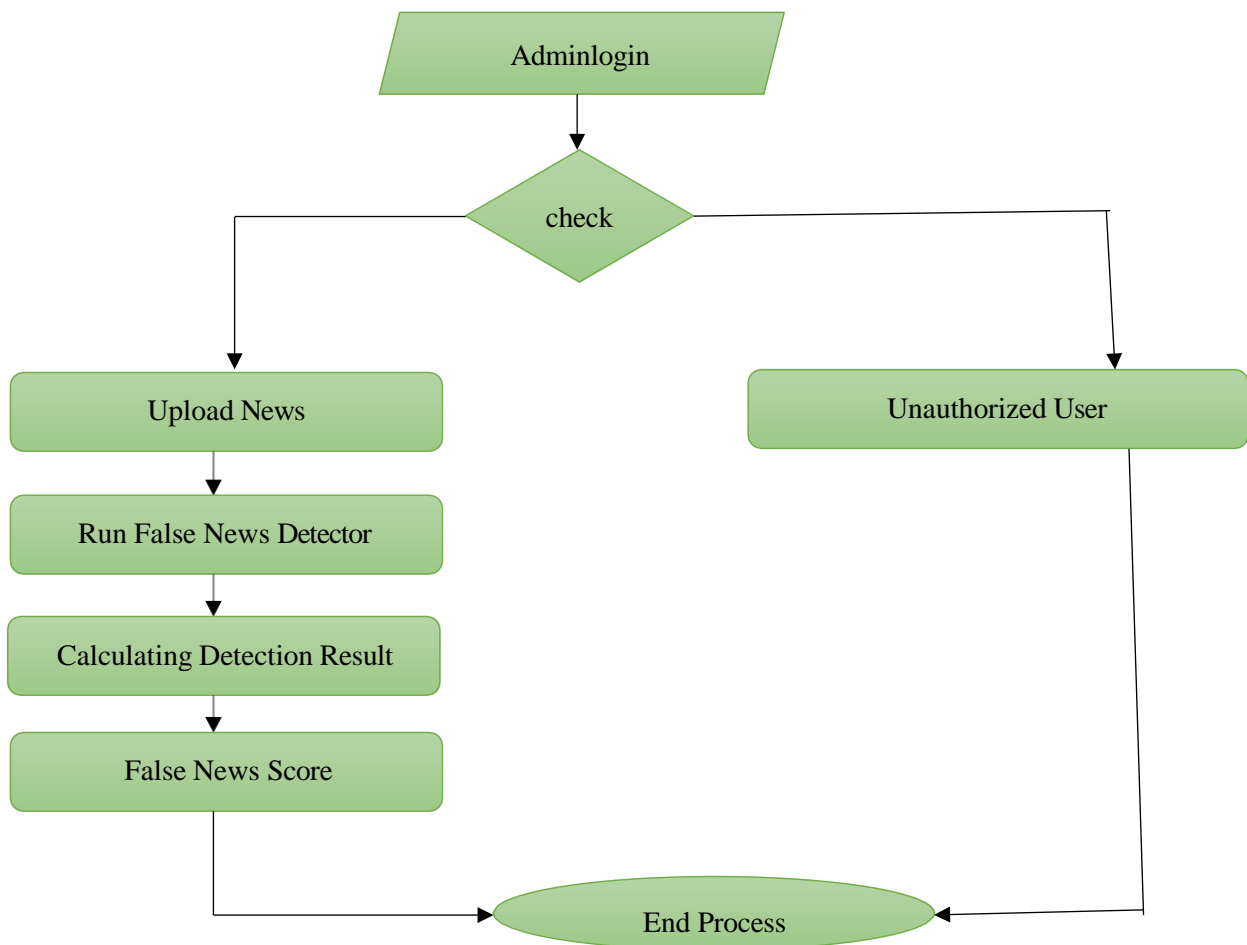


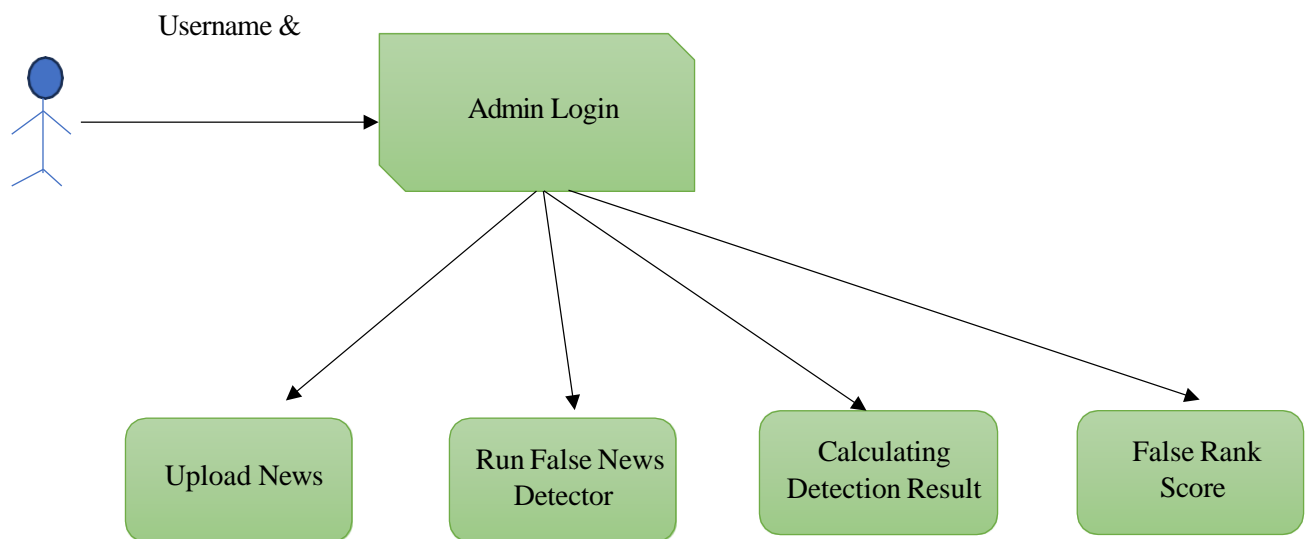
Figure 5.3.5 Data Flow Diagram

v. **COMPONENT DIAGRAM**

- Component diagram is a special kind of diagram in UML. The purpose is also different from all other diagrams discussed so far. It does not describe the functionality of the system but it describes the components used to make those functionalities.
- Thus from that point of view, component diagrams are used to visualize the physical components in a system. These components are libraries, packages, files, etc.
- Component diagrams can also be described as a static implementation view of a system. Static implementation represents the organization of the components at a particular moment.

- A single component diagram cannot represent the entire system but a collection of diagrams is used to represent the whole.

UML Component diagrams are used in modeling the physical aspects of object-oriented systems that are used for visualizing, specifying, and documenting component-based systems and also for constructing executable systems through forward and reverse engineering. Component diagrams are essentially class diagrams that focus on a system's components that often used to model the static implementation view of a system.



password

Figure 5.3.6 Component Diagram

vi. ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

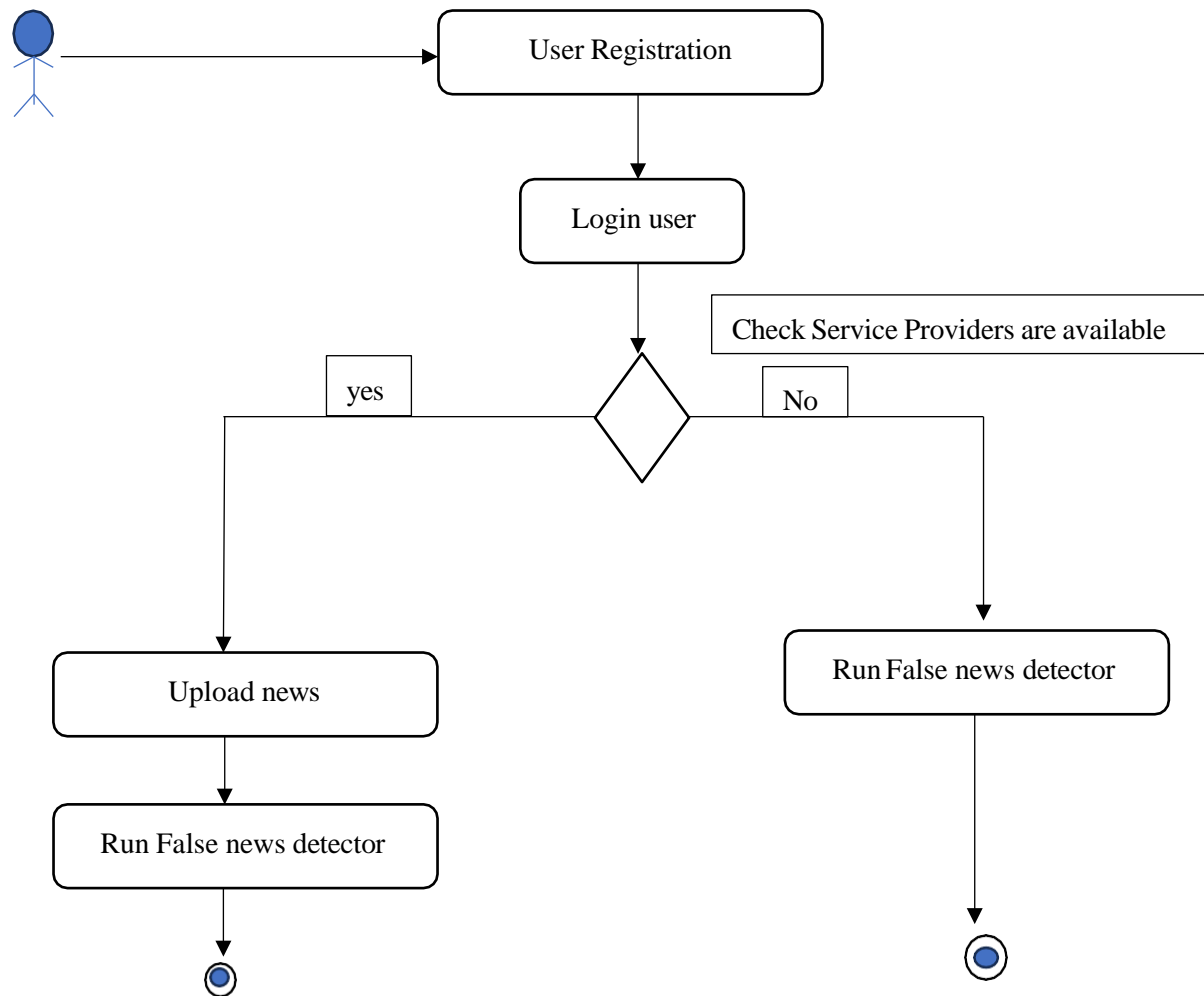


Figure 5.3.7 Activity Diagram

5.4 Summary:

The design of the Fake News Detection System follows a modular and systematic approach, combining Natural Language Processing (NLP) with a supervised learning algorithm to classify news articles as real or fake. The system is composed of several core components, each responsible for specific tasks.

CHAPTER-6

IMPLEMENTATION

AND RESULTS

6. IMPLEMENTATION AND RESULTS

6.1 Introduction :

In today's digital age, the implementation of the Fake News Detection System integrates Natural Language Processing (NLP) and Machine Learning (ML) techniques to classify news articles as Fake or Real. The purpose of this chapter is to provide a comprehensive overview of how the system was developed, the logical flow of processes, and the technologies involved in building a functional and efficient fake news classifier. The core idea behind the system is to identify linguistic features commonly associated with real news articles—such as the presence of quotes, named entities, and verbs—and use these features to calculate an attribution score. This score, along with supervised learning algorithms like Naive Bayes, is used to evaluate the authenticity of news content. The system is built using Python 3.7 and leverages powerful libraries such as TensorFlow, Scikit-learn, Pandas, and Matplotlib for data processing, feature extraction, machine learning, and visualization. The implementation follows a modular design to ensure scalability, reusability, and clarity.

The primary objectives of the implementation include:

- Developing a reliable NLP-based feature extractor.
- Designing a custom scoring function based on journalistic writing styles.
- Training a classifier to distinguish fake news from real news.
- Creating a user-friendly interface for uploading, analyzing, and reviewing results.

Modules:

The system architecture is divided into multiple functional modules, each responsible for a specific task in the Fake News Detection workflow. This modular approach improves scalability, maintainability, and clarity of the system. Below are the main modules used in the implementation:

1. User Interface (UI)
2. File Upload and Handling
3. Text Preprocessing
4. Feature Extraction
5. Attribution Scoring
6. Classification
7. Model Loading and Management
8. Visualization

6.2 Method of Implementation:

1. User Interface (UI) :

- **Description:** Manages user interaction and provides a graphical interface for input and output.
- **Functions:**
 - [1] Admin login verification
 - [2] File upload interface for .csv files
 - [3] Display of results (Fake/Real classification)

2. File Upload and Handling :

- **Description:** Handles file selection, upload, reading, and basic validation.
- **Functions:**
 - [1] Accept .csv input containing multiple news paragraphs
 - [2] Parse and prepare text content for processing
 - [3] Validate file format and content

3. Text Preprocessing Module :

- **Description:** Cleans and standardizes raw text for NLP processing

- **Functions:**

- [1] Convert text to lowercase
- [2] Remove punctuation and special characters
- [3] Tokenize sentences and words
- [4] Eliminate stop words

4. Feature Extraction :

- **Description:** Extracts meaningful linguistic features required for classification.
- **Techniques Used:**
 - [1] Named Entity Recognition (NER): Identifies names of persons, organizations, places.
 - [2] Verb Extraction: Extracts verbs using Part-of-Speech (POS) tagging.
 - [3] Quote Detection: Detects quoted statements using quotation marks and parsing rules.

5. Attribution Scoring :

- **Description:** Calculates a custom score that represents the journalistic quality of the article.
- **Formula:**
$$\text{Score} = \frac{\text{Number of Quotes} + \text{Named Entities} + \text{Verbs}}{\text{Total Sentence Length}}$$
- **Purpose:** Higher scores suggest well-sourced, credible journalism.

6. Classification:

- **Description:** Uses calculated score and machine learning to classify news as FALSE or REAL.
- **Technique Used:**
 - [1] Naive Bayes Classifier from Scikit-learn
 - [2] Threshold logic: $\text{Score} > 0.90 \rightarrow \text{REAL}$, else FAKE

7. Model Loading and Management :

- **Description:** Manages ML models used for inference.
- **Functions:**
 - [1] Load pre-trained models (e.g., MobileNet or custom-trained classifier)

[2] Handle inference pipeline

[3] Manage model files and logs

8. Visualization :

- **Description:** Displays classification results in a structured and user-friendly format.
- **Functions:**
 - [1] Show table of news text, classification result, and score
 - [2] Provide summary statistics for overall result

6.3 SAMPLE CODE:

```
1. from django.shortcuts import render
2. from django.shortcuts import render
3. from django.template import RequestContext
4. from django.contrib import messages
5. from django.http import HttpResponse
6. from django.conf import settings
7. from django.core.files.storage import FileSystemStorage
8. from textblob import TextBlob
9. import re
10. import nltk
11. global name
12. def index(request):
13.     if request.method == 'GET':
14.         return render(request, 'index.html', {})
15.     def Login(request):
16.         if request.method == 'GET':
17.             return render(request, 'Login.html', {})
18.     def UploadNews(request):
```

```
19. if request.method == 'GET':
20. return render(request, 'UploadNews.html', { })
21. def AdminLogin(request):
22. if request.method == 'POST':
23. username = request.POST.get('t1', False)
24. password = request.POST.get('t2', False)
25. if username == 'admin' and password == 'admin':
26. context= {'data': 'welcome '+username}
27. return render(request, 'AdminScreen.html', context)
28. else:
29. context= {'data': 'login failed'}
30. return render(request, 'Login.html', context)
31. def UploadNewsDocument(request):
32. global name
33. if request.method == 'POST' and request.FILES['t1']:
34. output = "
35. myfile = request.FILES['t1']
36. fs = FileSystemStorage()
37. name = str(myfile)
38. filename = fs.save(name, myfile)
39. context= {'data': name+' news document loaded'}
40. return render(request, 'UploadNews.html', context)
41. def getQuotes(paragraph): #checking paragraph contains quotes or not
42. score = 0
43. match = re.findall('(?"(.*)"')', paragraph)
44. if match:
45. score = len(match)
```

```
46.return score
47.def checkVerb(paragraph): #checking paragraph contains verbs or not
48.score = 0
49.b = TextBlob(paragraph)
50.list = b.tags
51.for i in range(len(list)):
52.arr = str(list[i]).split(",")
53.verb = arr[1].strip();
54.verb = verb[1:len(verb)-2]
55.if verb == 'VBG' or verb == 'VBN' or verb == 'VBP' or verb == 'VBD':
56.score = score + 1
57.return score
58.def nameEntities(paragraph): #getting names from paragraphs
59.score = 0
60.for chunk in nltk.ne_chunk(nltk.pos_tag (nltk.word_tokenize(paragraph))):
61.if hasattr(chunk, 'label'):
62.name = ''.join(c[0] for c in chunk)
63.score = score + 1
64.return score
65.def naiveBayes(quotes_score, verb_score, name, paragraph): #Naive Bayes to
    calculate score
66.score = quotes_score + verb_score + name
67.arr = nltk.word_tokenize(paragraph)
68.total = (score/len(arr) * 10)
69.return total
70.def DetectorAlgorithm(request): #detector and classifier algorithm
```

```
71. global name
72. if request.method == 'GET':
73. strdata = '<table border=1 align=center width=100%><tr><th>News
    Text</th><th>ClassifierDetectionResult</th><th>FakeRankScore
    </th></tr><tr>'
74. with open(name, "r") as file:
75. for line in file:
76. line = line.strip('\n')
77. line = line.strip()
78. quotes_score = getQuotes(line)
79. verb_score = checkVerb(line)
80. entity_name = nameEntities(line)
81. score = naiveBayes(quotes_score, verb_score, entity_name, line)
82. if score > 0.90:
83. strdata+='<td>'+line+'</td><td>Real News</td><td>'+str(score)+'</td></tr>'
84. else:
85. strdata+='<td>'+line+'</td><td>false News</td><td>'+str(score)+'</td></tr>'
86. context= {'data':strdata}
87. return render(request, 'ViewFakeNewsDetector.html', context)
```

6.4 OUTPUT SCREENS:

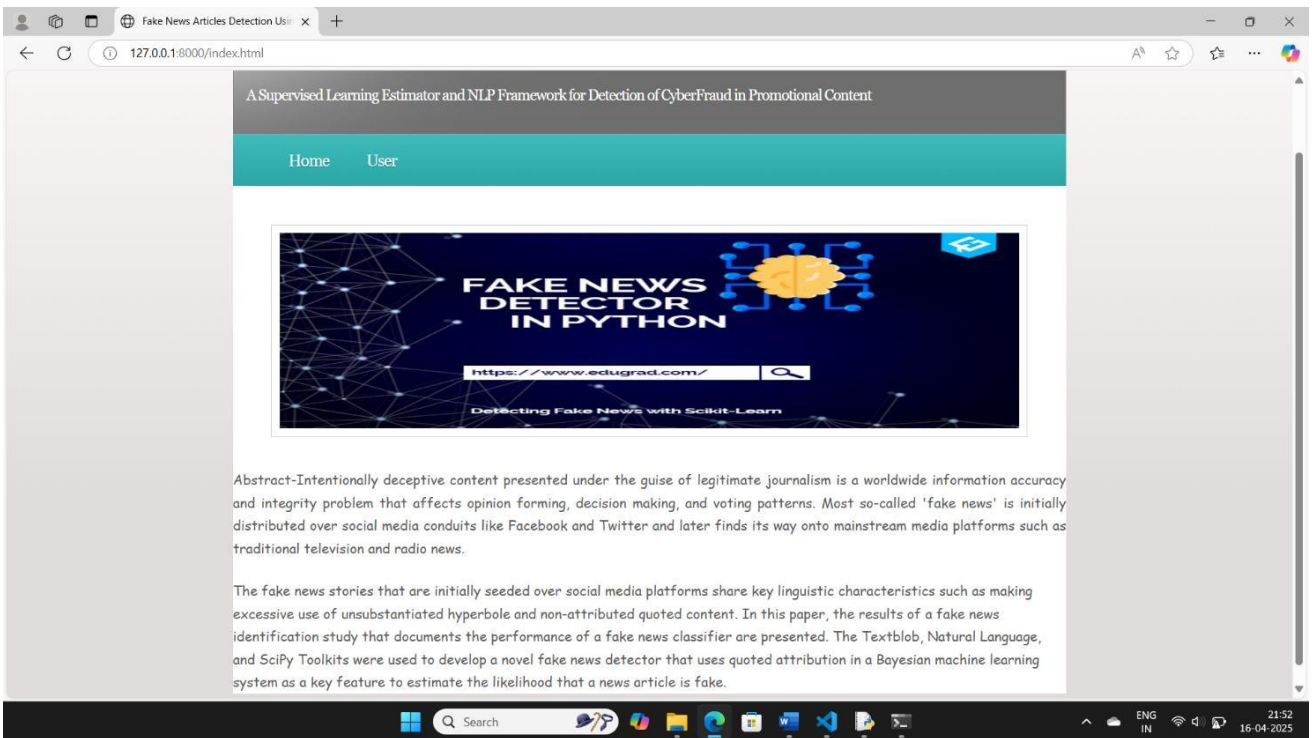


Figure 6.4.1. Home Page

In above screen click on 'User' link to get below screen

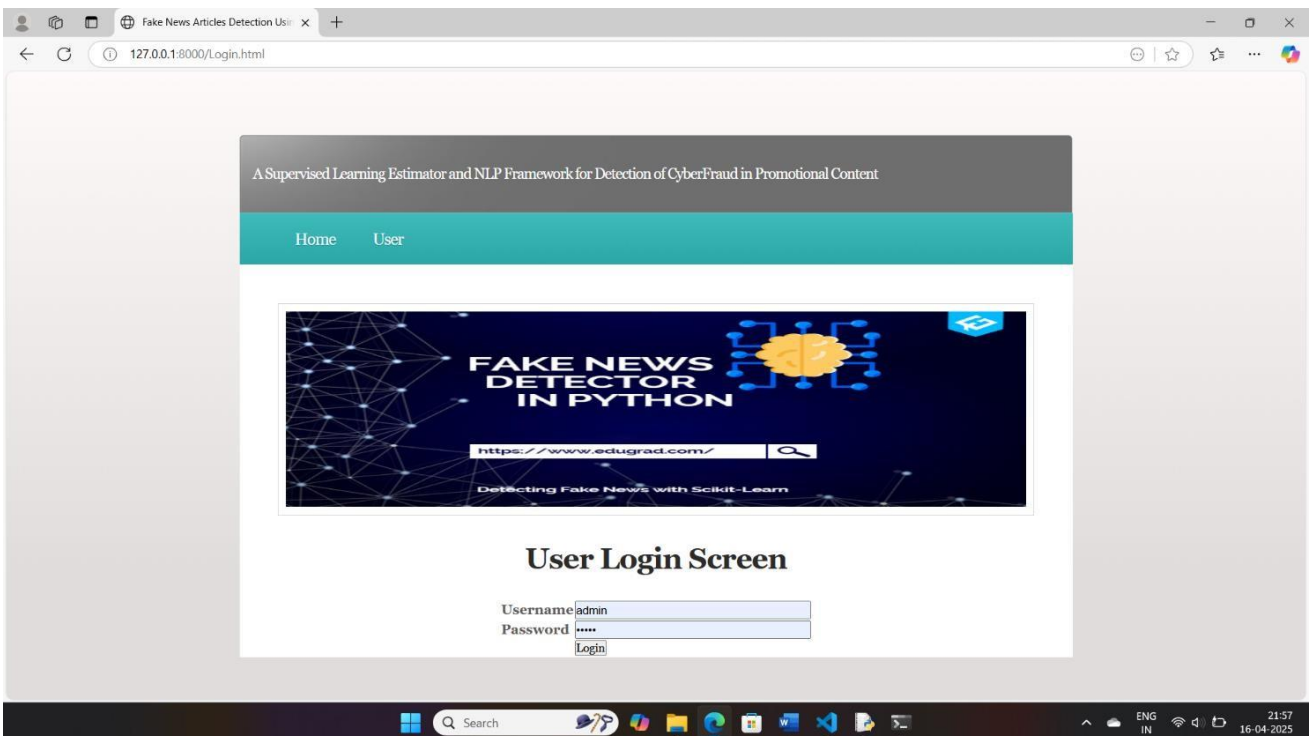


Figure 6.4.2 User Login Page

In above screen enter username and password as 'admin' and then click on 'Login' button to get below screen

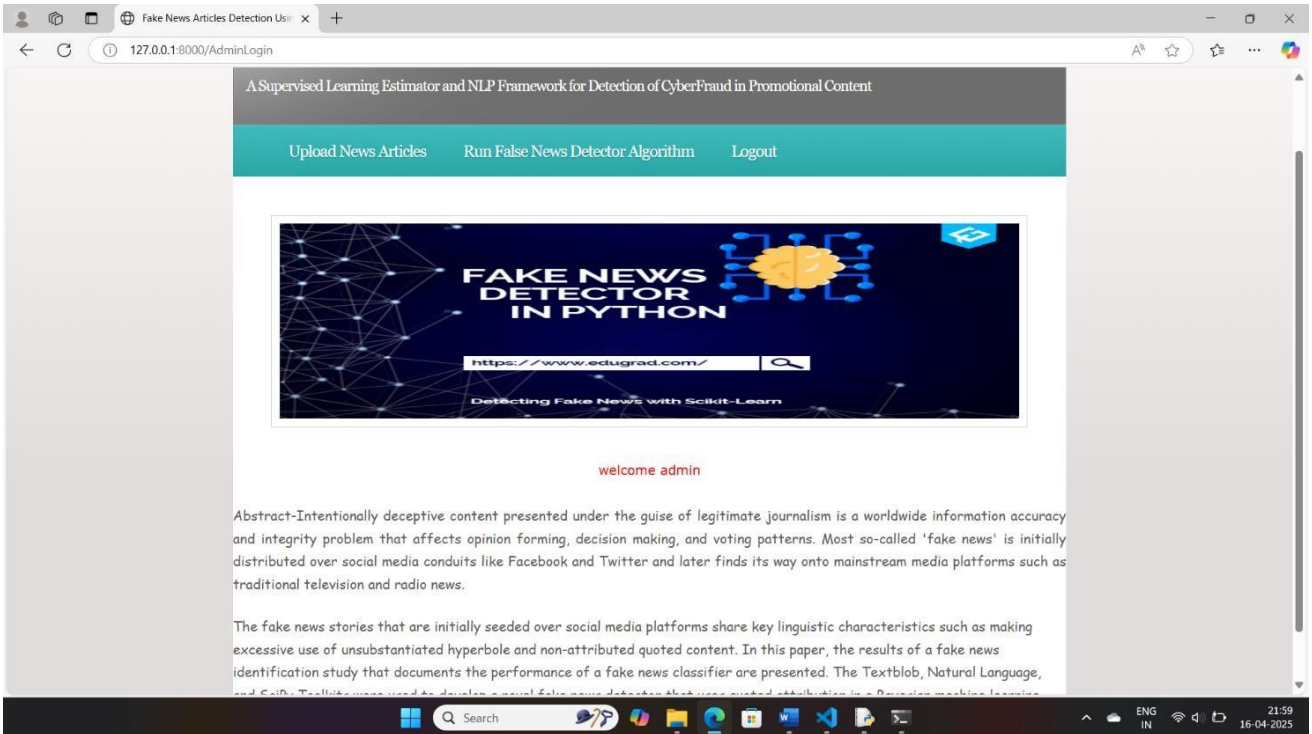


Figure 6.4.3. Upload News Article Page

In above screen click on 'Upload News Articles' link to upload news document

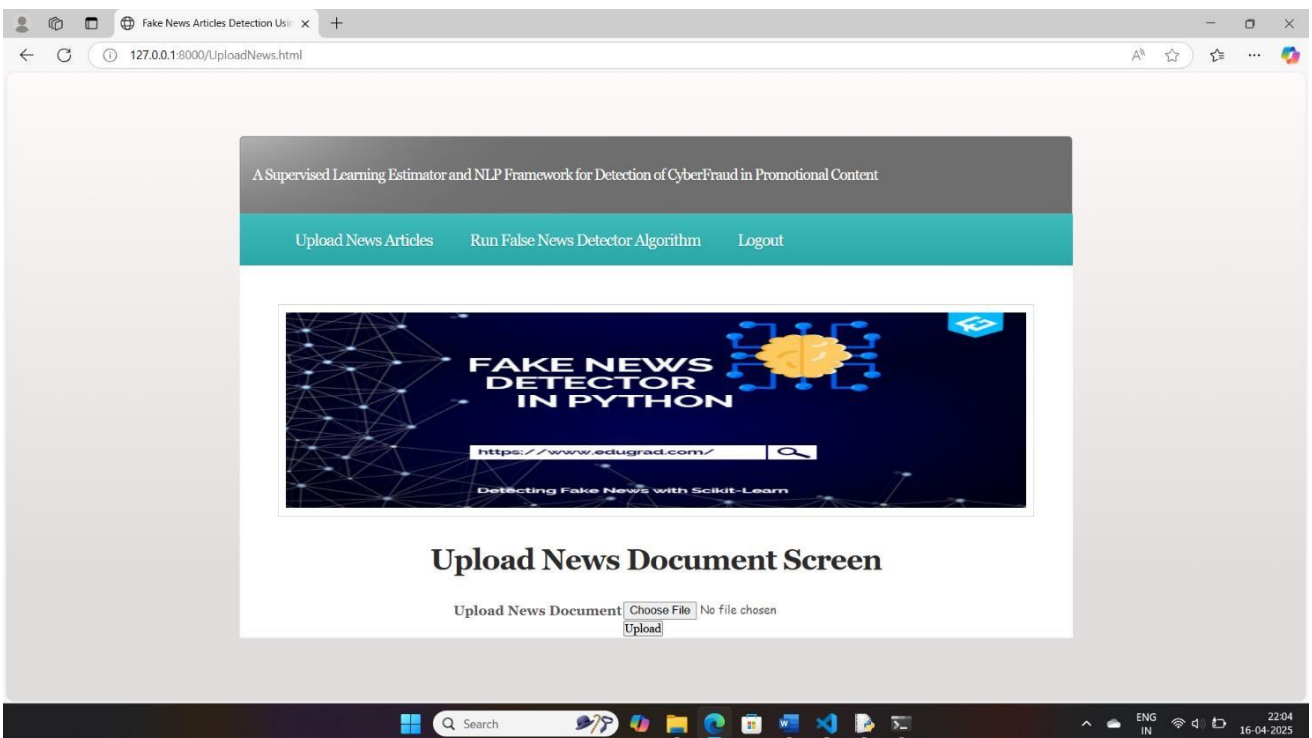


Figure 6.4.4 Upload News Document Page

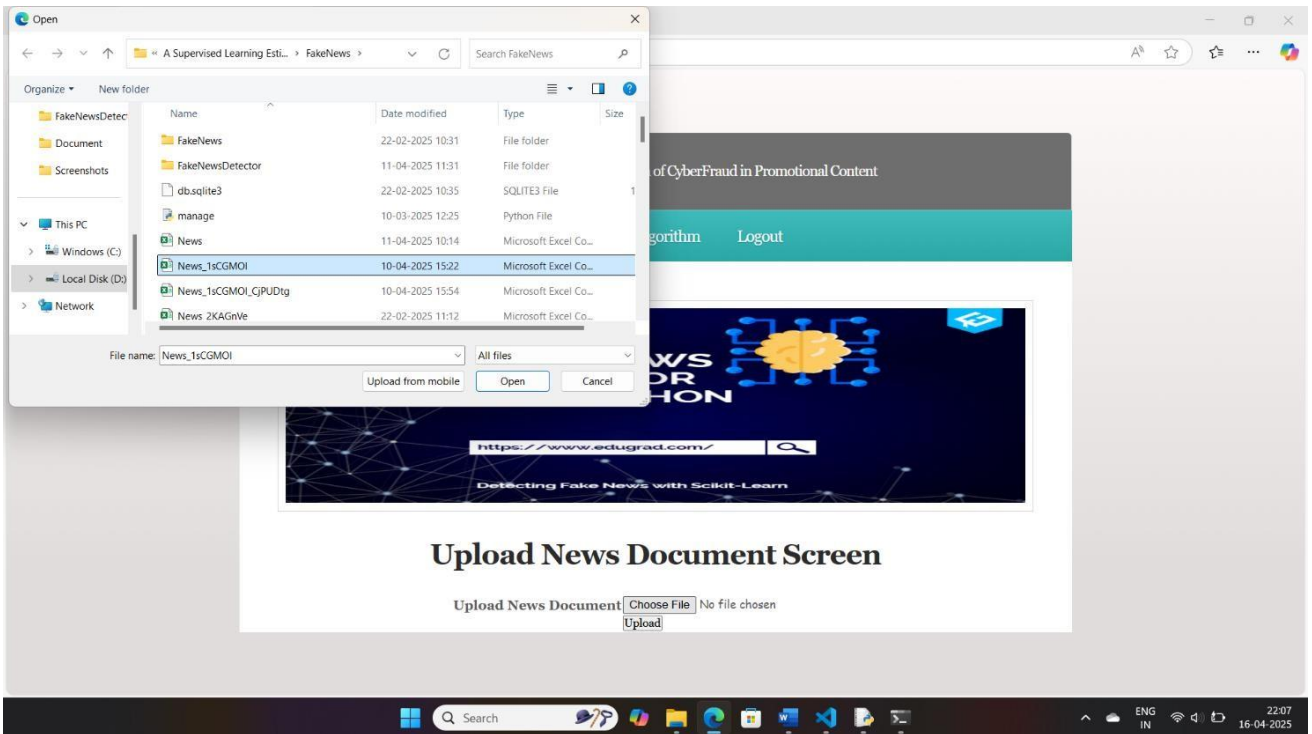


Figure 6.4.5 Uploading Dataset

In above screen I am uploading 'News.csv' file which contains 150 news paragraphs. After uploading news will get below screen

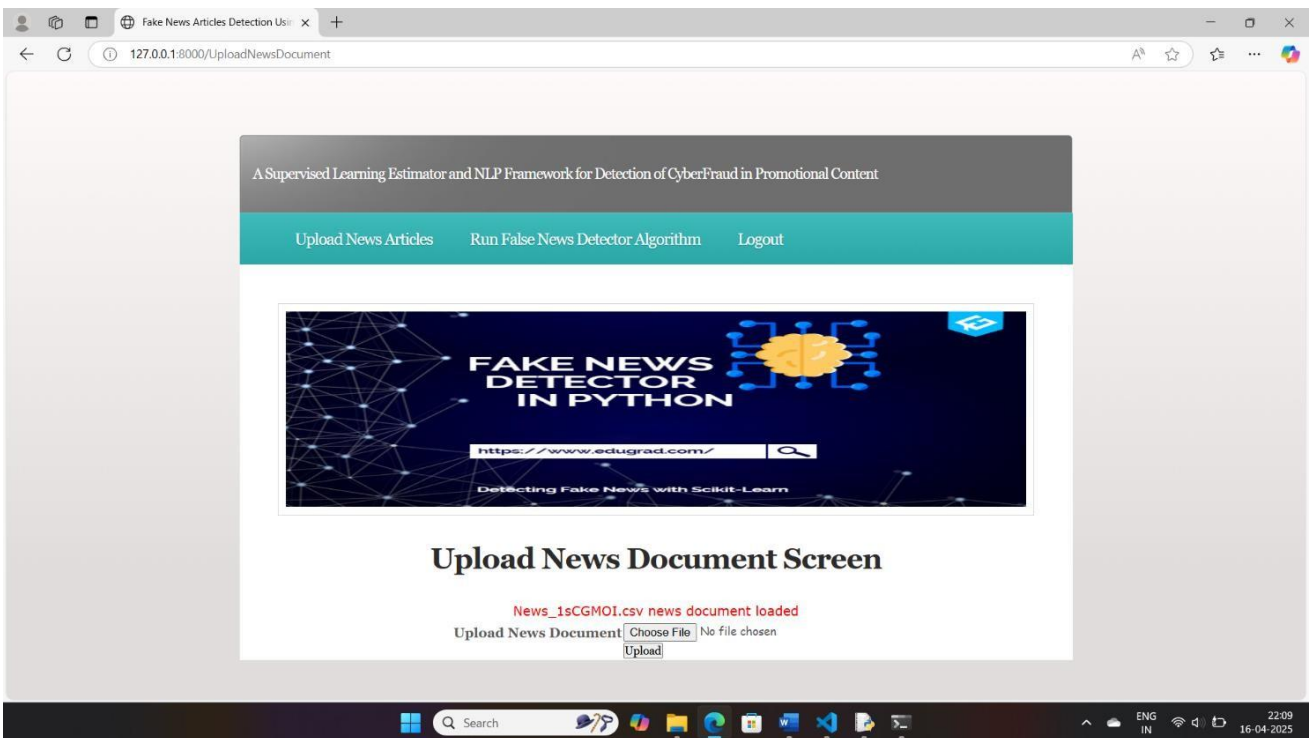
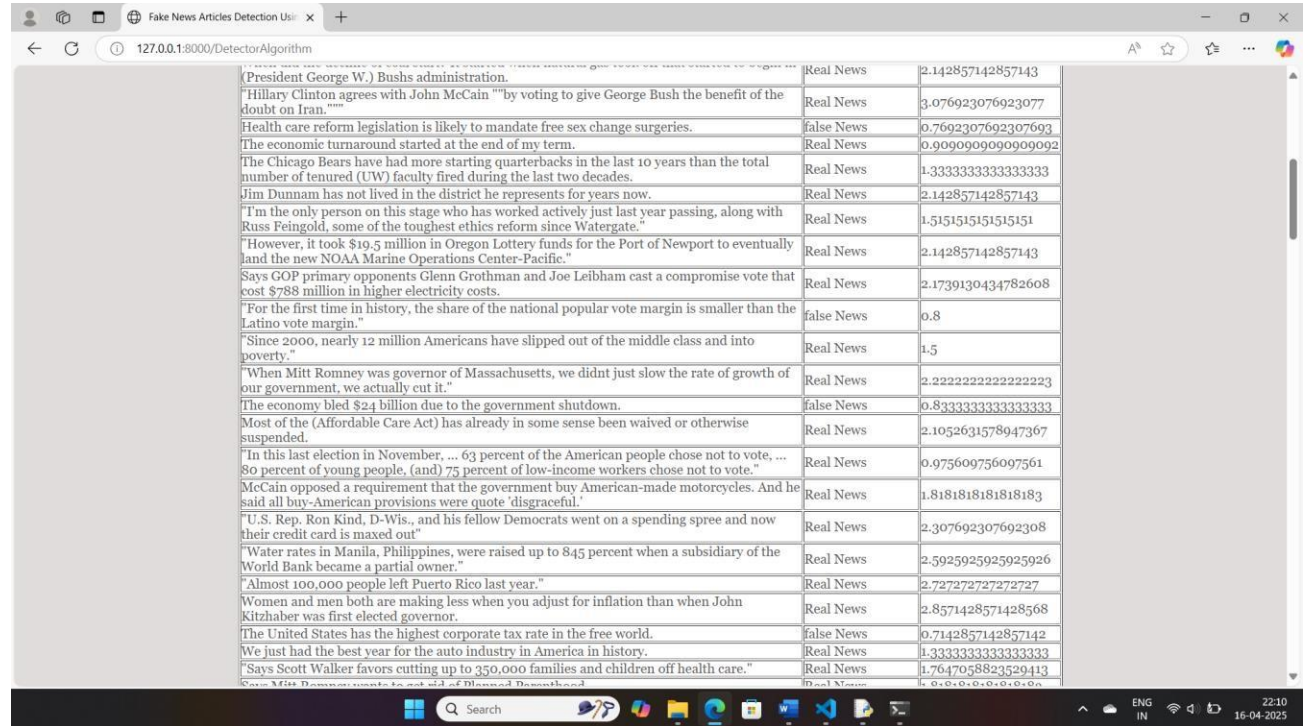


Figure 6.4.6 After Uploading Dataset Page

In above screen news file uploaded successfully, now click on 'Run Fake News Detector Algorithm' link to calculate Fake News Detection algorithm score and based on score and naïve bayes algorithm we will get result.



(President George W.) Bush's administration.	Real News	2.142857142857143
"Hillary Clinton agrees with John McCain ""by voting to give George Bush the benefit of the doubt on Iran.""	Real News	3.076923076923077
Health care reform legislation is likely to mandate free sex change surgeries.	false News	0.7692307692307693
The economic turnaround started at the end of my term.	Real News	0.9090909090909092
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	Real News	1.3333333333333333
Jim Dunnam has not lived in the district he represents for years now.	Real News	2.142857142857143
"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."	Real News	1.5151515151515151
"However, it took \$19.5 million in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific."	Real News	2.142857142857143
Says GOP primary opponents Glenn Grothman and Joe Leibham cast a compromise vote that cost \$788 million in higher electricity costs.	Real News	2.1739130434782608
"For the first time in history, the share of the national popular vote margin is smaller than the Latino vote margin."	false News	0.8
"Since 2000, nearly 12 million Americans have slipped out of the middle class and into poverty."	Real News	1.5
"When Mitt Romney was governor of Massachusetts, we didnt just slow the rate of growth of our government, we actually cut it."	Real News	2.2222222222222223
The economy bled \$24 billion due to the government shutdown.	false News	0.8333333333333333
Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended.	Real News	2.1052631578947367
"In this last election in November, ... 63 percent of the American people chose not to vote, ... 80 percent of young people, (and) 75 percent of low-income workers chose not to vote."	Real News	0.975609756097561
McCain opposed a requirement that the government buy American-made motorcycles. And he said all buy-American provisions were quote 'disgraceful.'	Real News	1.8181818181818183
"U.S. Rep. Ron Kind, D-Wis., and his fellow Democrats went on a spending spree and now their credit card is maxed out"	Real News	2.307692307692308
"Water rates in Manila, Philippines, were raised up to 845 percent when a subsidiary of the World Bank became a partial owner."	Real News	2.5925925925925926
"Almost 100,000 people left Puerto Rico last year."	Real News	2.727272727272727
Women and men both are making less when you adjust for inflation than when John Kitzhaber was first elected governor.	Real News	2.8571428571428568
The United States has the highest corporate tax rate in the free world.	false News	0.7142857142857142
We just had the best year for the auto industry in America in history.	Real News	1.3333333333333333
"Says Scott Walker favors cutting up to 350,000 families and children off health care."	Real News	1.7647058823529413

Figure 6.4.7 Showing Either FALSE or REAL News

In above screen first column contains news text and second column is the result value as 'fake or real' and third column contains score. If score greater > 0.90 then I am considering news as REAL otherwise fake.

- For all 150 news text articles we got result as fake or real.
- Above screen shots of code calculating quotes, name entity and verbs from news paragraphs

CHAPTER-7

TESTING AND VALIDATION

7. TESTING AND VALIDATION

7.1 Introduction:

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, subassemblies, 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.

7.2 Types of Tests:

7.2.1 Unit 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.

7.2.2 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 satisfactory, 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.

7.2.3 Functional Test :

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. Functions : identified functions must be exercised. Output : identified classes of application outputs must be exercised. Systems/Procedures: interfacing systems or procedures must be invoked.

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.

7.2.4 System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.2.5 White 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.

7.2.6 Black Box Testing :

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.

7.2.7 Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.2.8 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.

Test Results: All the test cases mentioned above passed successfully. No defects encountered

CHAPTER-8

CONCLUSION

8. CONCLUSION

The paper presented the results of a study that produced a limited false news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

CHAPTER-9

REFERENCES

9. REFERENCES

- [1] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing," in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 9–18.
- [2] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "Gps estimation for places of interest from social users' uploaded photos," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 2058–2071, 2013.
- [3] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized poi recommendation," IEEE Transactions on Multimedia, vol. 17, no. 6, pp. 907–918, 2015.
- [4] J. Sang, T. Mei, and C. Sun, J.T.and Xu, "Probabilistic sequential pois recommendation via check-in data," in Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2012.
- [5] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," ACM Transactions on the Web, vol. 5, no. 1, p. 5, 2011.
- [6] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in Proceedings of 29th International Conference on AAAI. AAAI, 2015.

- [7] Q. Yuan, G. Cong, and A. Sun, “Graph-based point-of-interest recommendation with geographical and temporal influences,” in Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. ACM, 2014, pp. 659–668.
- [8] H. Yin, C. Wang, N. Yu, and L. Zhang, “Trip mining and recommendation from geo-tagged photos,” in IEEE International Conference on Multimedia and Expo Workshops. IEEE, 2012, pp. 540–545.
- [9] Y. Gao, J. Tang, R. Hong, Q. Dai, T. Chua, and R. Jain, “W2go: a travel guidance system by automatic landmark ranking,” in Proceedings of the international conference on Multimedia. ACM, 2010, pp. 123–132.
- [10] X. Qian, Y. Zhao, and J. Han, “Image location estimation by salient region matching,” IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4348–4358, 2015.
- [11] H. Kori, S. Hattori, T. Tezuka, and K. Tanaka, “Automatic generation of multimedia tour guide from local blogs,” Advances in Multimedia Modeling, pp. 690–699, 2006.
- [12] T. Kurashima, T. Tezuka, and K. Tanaka, “Mining and visualizing local experiences from blog entries,” in Database and Expert Systems Applications. Springer, 2006, pp. 213–222.

- [13] Y. Shi, P. Serdyukov, A. Hanjalic, and M. Larson, “Personalized landmark recommendation based on geo-tags from photo sharing sites,” ICWSM, vol. 11, pp. 622–625, 2011.
- [14] M. Clements, P. Serdyukov, A. de Vries, and M. Reinders, “Personalised travel recommendation based on location co-occurrence,” arXiv preprint arXiv:1106.5213, 2011.
- [15] X. Lu, C. Wang, J. Yang, Y. Pang, and L. Zhang, “Photo2trip: generating travel routes from geo-tagged photos for trip planning,” in Proceedings of the international conference on Multimedia. ACM, 2010, pp. 143–152.
- [16] Y. Zheng, L. Zhang, X. Xie, and W. Ma, “Mining interesting locations and travel sequences from gps trajectories,” in Proceedings of the 18th international conference on World wide web. ACM, 2009, pp. 791–800.
- [17] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, “Collaborative location and activity recommendations with gps history data,” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 1029–1038.
- [18] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, “Discovering urban functional zones using latent activity trajectories,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 712–725, 2015. [Online]. Available: <http://dx.doi.org/10.1109/TKDE.2014.2345405>

- [19] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan, “Discovering areas of interest with geo-tagged images and check-ins,” in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 589–598.
- [20] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang, “Summarizing tourist destinations by mining user-generated travelogues and photos,” *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 352–363, 2011.