**PROJECT REPORT**

**ON**

**WHAT’S APP NEWS**

**Submitted for Partial Fulfilment of Award of**

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE & ENGINEERING**

**(2019-2023)**

**By**

**ABHISHEK KUMAR SINGH (1903630100007)**

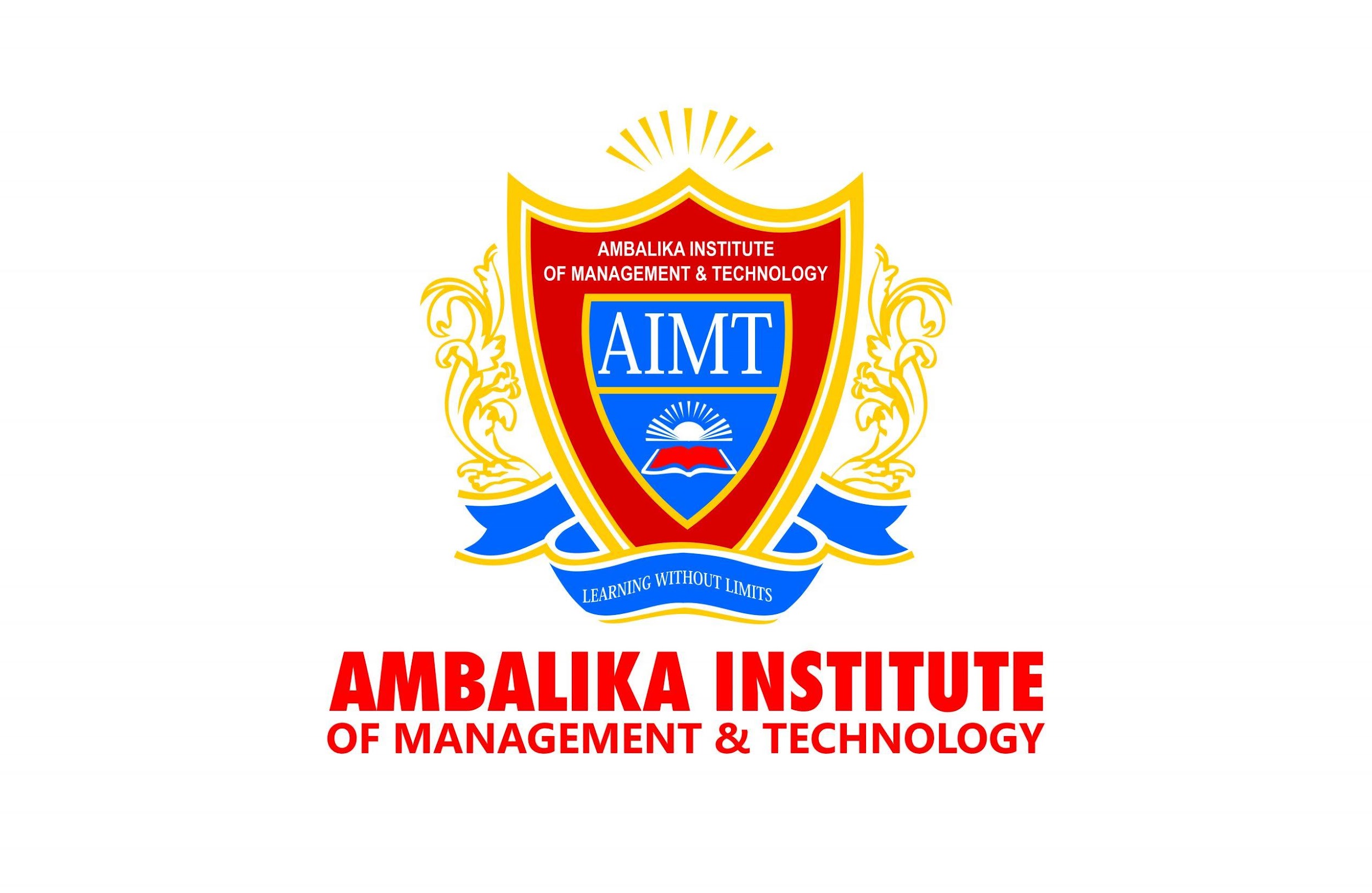
**ASHEESH KUMAR(1903630100028)**

**KARTIKAY MISHRA(1903630100053)**

**Under the Guidance**

**Of**

**Mr. VIPIN RAWAT**



**Submitted to**

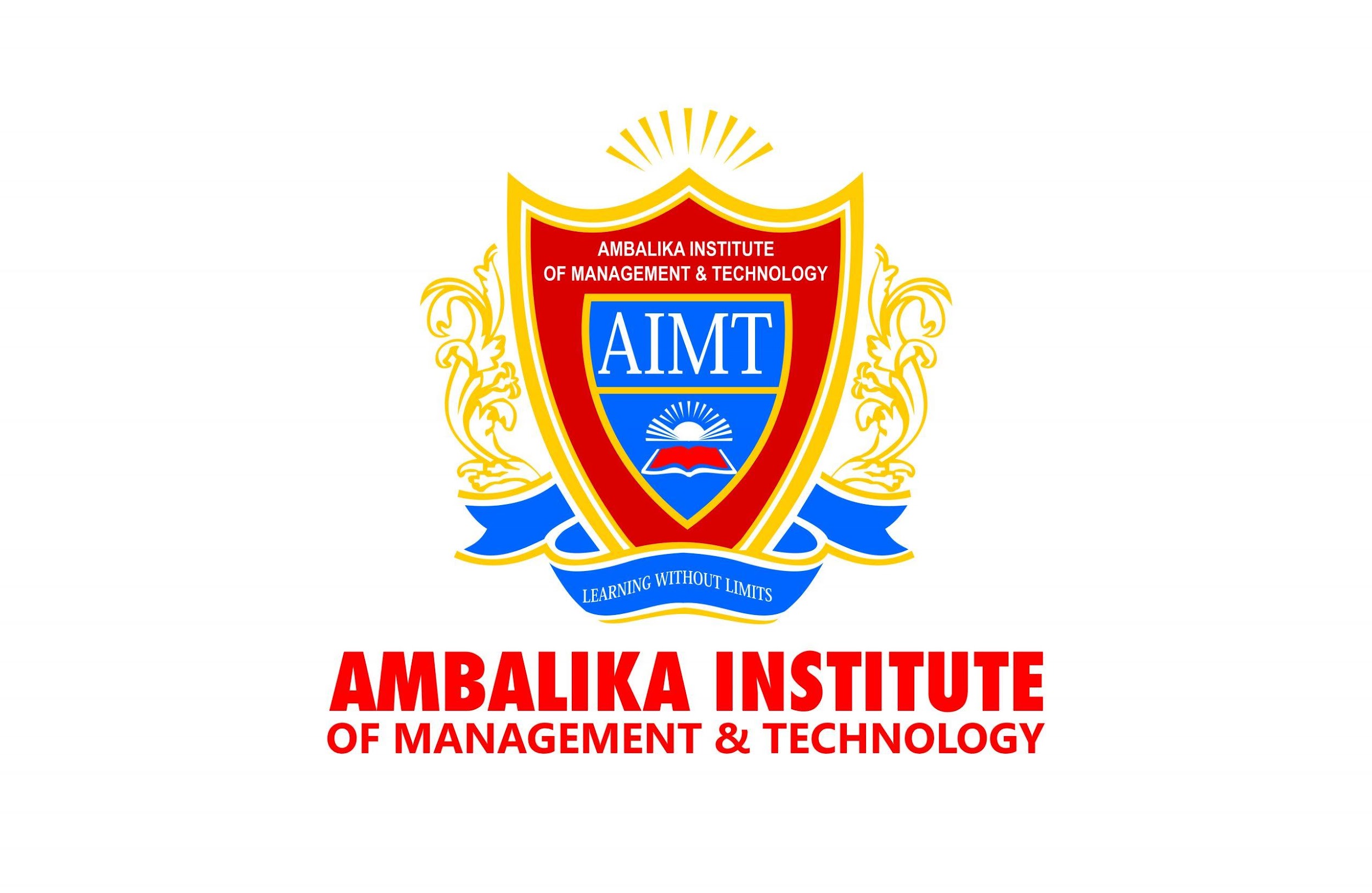
**AMBALIKA INSTITUTE OF MANAGEMENT & TECHNOLOGY, LUCKNOW**

**Affiliated to**

**DR**. **A**. **P**. **J**. **ABDUL KALAM TECHNICAL UNIVERSITY**, **LUCKNOW**, **INDIA**.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

**AMBALIKA INSTITUTE OF MANAGEMENT & TECHNOLOGY, LUCKNOW**



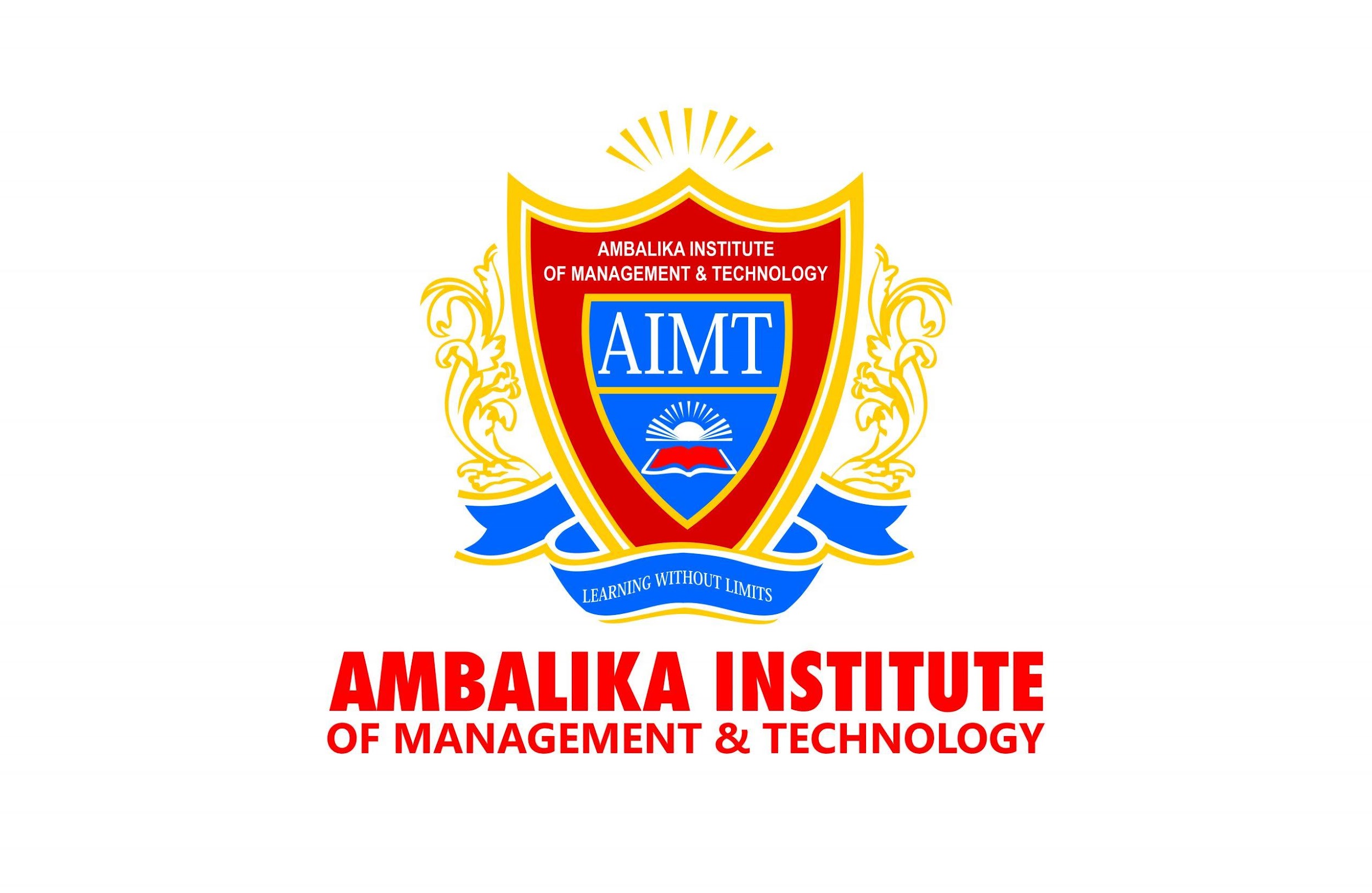
**CERTIFICATE**

Certified that the Project entitled “WHATSAPP NEWS” by ABHISHEK KUMAR SINGH (1903630100007) ASHEESH KUMAR (190363100028) KARTIKAY MISHRA (1903630100053) in the partial fulfillment of the requirements for the award of the degree of Bachelor Of Technology in Computer Science & Engineering of Dr. A. P. J. ABDUL KALAM TECHNICAL UNIVERSITY, LUCKNOW, INDIA is a record of her/his work carried under my supervision and guidance. The project report embodies the result of original work and activities carried out by his/her and the contents of the project report do not form the basis for the award of my other degree to the candidate or anybody else.

**Mr. VIPIN RAWAT Mr. VINAY KUMAR**

**Asst. Prof. CSE HOD CSE**

**(PROJECT GUIDE)**



**Department of Computer Science & Engineering**

**DECLARATION**

We hereby declare that the project “WHATSAPP NEWS” submitted by us in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science & Engineering of DR. A. P. J. ABDUL KALAM TECHNICAL UNIVERSITY, LUCKNOW, is a record of our work carried under the supervision and guidance of Mr. VINAY KUMAR. To the best of my knowledge, the project has been submitted to DR. A. P. J. KALAM TECHNICAL UNIVERSITY, LUCKNOW, or any other University or Institute for the award of any degree.

**ABHISHEK KUMAR SINGH KARTIKAY MISHRA**

**(1903630100007) (1903630100053)**

**ASHEESH KUMAR**

**(1903630100028)**

**ACKNOWLEDGEMENT**

This is my proud privilege in expressing a deep sense of obligation and gratitude to “Mr. VIPIN RAWAT” who has assigned us our project on such a rising and interesting topic, which is the “WHATSAPP NEWS”. We are also thankful to him for his support and valuable guidance rejuvenating encouragement, positive criticism, and constant supervision throughout our project.

We see greatly indebted to be associated with him for his everlasting impression, scientific temperament, and humanitarian sensibility. He has always been seen as a source of inspiration to us throughout the work.

We express sincere and intense thanks to our HOD “MR. VINAY KUMAR” and all the computer science and engineering department members for arranging these kinds of projects for the students.

We also express our gratitude to our college for providing us with the infrastructure to carry out the project and to all staff members who were directly or indirectly instrumental in enabling us to stay committed to the project.

**ABHISHEK KUMAR SINGH KARTIKAY MISHRA**

**(1903630100007) (1903630100053)**

**ASHEESH KUMAR**

**(1903630100028)**

**ABSTRACT**

This application is designed to automatically scrape news from a news portal and deliver it to users via WhatsApp. The application utilizes web scraping techniques to extract news articles and their related information from a specified news portal. The extracted data is then processed and formatted into a format that is suitable for sharing on WhatsApp.

The application is designed to be highly customizable, allowing users to select their preferred news portal, topics, and frequency of updates. Users can specify the types of news they want to receive, such as international news, local news, business news, sports news, etc.

The application is also designed to be highly automated, requiring minimal user intervention. Once the user has configured the settings, the application will automatically scrape the news portal at specified intervals and send the news updates directly to the user's WhatsApp account. The user can then read the news updates and share them with their contacts.

Overall, this application provides an easy and convenient way for users to stay up-to-date on the latest news from their preferred news portal, without the need for constantly checking different websites or apps. By automating the process, users can receive news updates in real-time, directly on their mobile devices.

INDEX

**CHAPTER # 1**

**INTRODUCTION**

**Project Introduction:-**

WhatsApp is a popular messaging application that has become an increasingly important platform for news distribution. With over 2 billion active users worldwide, it is now one of the largest social media platforms in the world. WhatsApp allows users to share news articles, videos, and other types of media with their contacts, making it a powerful tool for news dissemination.

The use of WhatsApp for news distribution has grown rapidly in recent years, particularly in developing countries where it is widely used. However, the platform has also faced challenges, including the spread of fake news and misinformation, which has led to incidents of violence and social unrest in some countries.

In response to these challenges, WhatsApp has implemented measures to combat the spread of misinformation on its platform, such as limiting the number of times a message can be forwarded, labeling forwarded messages, and working with fact-checking organizations to verify the accuracy of news stories.

Despite these challenges, WhatsApp remains an important platform for news distribution, particularly in regions with limited access to traditional news sources. With its large user base and powerful features, WhatsApp has the potential to shape the way news is shared and consumed in the years to come.

**Background:-**

WhatsApp is a cross-platform instant messaging app that was founded in 2009 by Brian Acton and Jan Koum. The app became very popular and was eventually acquired by Facebook in 2014 for $19 billion. Today, WhatsApp is one of the most popular messaging apps in the world, with over 2 billion active users.

The history of news on WhatsApp can be traced back to its early days when it was primarily used for personal messaging. However, as the app grew in popularity, people began to use it for group messaging and sharing information with larger audiences. This led to the emergence of WhatsApp groups, which are essentially chat rooms that can accommodate up to 256 people.

WhatsApp groups quickly became a popular way for people to share news and information with their friends and family. However, the app was not designed as a news platform, and there were concerns about the spread of fake news and misinformation. In 2018, WhatsApp introduced a feature that limited the forwarding of messages to only five contacts at a time, in an attempt to curb the spread of false information.

Despite these efforts, WhatsApp has continued to be used as a platform for sharing news and information. Many news organizations and journalists now use WhatsApp to distribute news and engage with their audiences. For example, the BBC has a WhatsApp service that sends subscribers news alerts and updates, while journalists in countries like India and Brazil use the app to connect with sources and share breaking news.

Overall, the history of news on WhatsApp is closely tied to the evolution of the app itself. While it was not designed as a news platform, WhatsApp has become an important tool for sharing information and staying connected with others. As the app continues to grow and evolve, news and information sharing will likely remain a key part of its functionality.

**Problem definition:-**

* the traditional system of delivering newspapers was time-consuming, labor-intensive, and also not so environmentally friendly.
* Some people use to read more than one paper daily to get overall unbiased news from the papers which cost them a lot and also requires more paper for printing.
* Other people only like to read a specific segment of the newspaper which in return has wasted the paper in manufacturing the paper and printing inks.
* Producing and distributing paper-based newspapers can be expensive, particularly for smaller newspapers with limited resources. These costs can be further compounded by declining readership, which can make it difficult for newspapers to generate enough revenue to cover production and distribution costs.
* Paper-based newspapers are limited in their reach, as they can only be distributed in specific geographic areas. This can make it difficult for newspapers to reach readers in other regions or countries.
* The production of paper-based newspapers can have a significant environmental impact, particularly in terms of deforestation and carbon emissions. As consumers become more environmentally conscious, this can be a deterrent for some readers who are looking for more sustainable news sources
* Paper-based newspapers are static, one-way communication channels that offer limited interactivity compared to digital news sources. This can make it difficult for readers to engage with the news and participate in conversations about current events.

Aim and objective of the project:-

* to develop and implement a web-based application with storage to automate the process of accessing the news from various news sites.
* To move the traditional newspaper reader to this WhatsApp-based system as the application is very simple to use and does not require a lot of interaction from the user.

**System Analysis:-**

**Introduction:**Analysis can be defined as breaking up any hole to find out its nature, function, etc. It defines design as making preliminary sketches; to sketch up a pattern or outline for a plan. To plan and carry out, especially by the artistic arrangement or in a skillful wall. System analysis and design can be characterized as a set of techniques and processes, a community of interest, a culture, and an intellectual orientation The various tasks in the system analysis include the following: -

* Understanding Application
* Planning
* Scheduling
* Performing cost-benefit analysis
* Supervising, Installing, and Maintaining the system

**Feasibility Study:-**

Feasibility Analysis begins once the goals are defined. It starts by generating broad possible solutions, which are possible to indicate what the new system should look like this is where creativity and imagination are used. Analysts must think up new ways of doing things-generate new ideas. There is no need to go into the detailed system operations yet. The solution should provide enough information to make reasonable estimates about project costs and give users an indication of how the new system will fit into the organization. It is important not to exert considerable effort at this stage only to find out that the project is not worthwhile or that there is a need to significantly change the original goal. Feasibility of a new system means ensuring that the new system which we are going to implement, is efficient and affordable.

The very first phase in any system developing life cycle is a preliminary investigation. In the preliminary study, we examine the project feasibility.

This project has been tested in the following areas of feasibility: -

 **Operational Feasibility**

 **Technical Feasibility**

 **Economic Feasibility**

**Operational Feasibility:-**

It was decided that the proposed system could be created as a web-based system that is operated by all the clients.

 The system is designed in such a way that it is easy to operate.

 The system does not involve any complex operations. It handles only simple functions.

 It is aimed at giving an easy way to generate news reports.

 The functional accessibility is also structured in such a way it won’t take much time for anyone to get accustomed to the system.

 The amount of knowledge and expertise required for being able to use this system is not much more than what it requires to operate a computer system.

 The client must have only a basic knowledge of web browsers.

Hence, there is no need for any special person with any particular qualification or expertise in any field to use WhatsApp News.

**Technical Feasibility:**

It was decided that WhatsApp News was technically feasible because of the following-

 The system has a very simple structure and an easy-to-understand body.

 Also, the software used to build the system is simple enough in its usage and functionality.

 To operate this system only a web browser is needed. The system is platform-independent but browser dependent as it requires a Chrome driver to bypass the sign-up using the cookies stored.

 The system can be expended if so decided.

 This system is technically more secure.

**Economic Feasibility:**

The Economic Feasibility of the system is mainly concerned with its financial aspects. It determines whether the investment that goes into the implementation of the project is recoverable or not. The Cost Benefit Analysis is a commonly used method in evaluating the effectiveness of the system.

 The cost of developing the system currently includes the minimum software cost to develop and run the system.

 Most of the software used in the system development is freely available on the web.

**Scope of Project:-**

 The project focuses on Web application development.

 The project is built around Epaper as the current system for its user that is still relying on traditional methods for consuming paper.

**CHAPTER # 2**

**Hardware and Software Requirements**

**Contents:-**

▪ Introduction

▪ System environment

▪ Software requirement

▪ Hardware requirements

**Introduction:-**

In this chapter, we mentioned the software and hardware requirements, which are necessary for successfully running this system. The major element in building systems is selecting compatible hardware and software. The system analyst has to determine what software package is best for “WhatsApp News” and, where software is not an issue, the kind of hardware and peripherals needed for the final conversion.

**System Environment:-**

After analysis, some resources are required to convert the abstract system into the real one. The hardware and software selection begins with requirement analysis, followed by a request for proposal and vendor evaluation. Software and real system are identified. According to the provided functional specification, all the technologies and their capacities are identified. Basic functions and procedures and methodologies are prepared for implementation. Some of the Basic requirements such as hardware and software are described as follows: -

**Hardware and Software Specification:-**

**Software Requirements:-**

* Technology: Web Scrapping with Automation
* IDE : Pycharm/Visual Studio Code
* Client Side Technologies: WhatsApp Web
* Server-Side Technologies: Python
* Operating System: Microsoft Windows/Linux
* Automation: Selenium
* PDF: PyPDF2
* Browser: Google Chrome

**Hardware Requirements:-**

* Processor: Pentium-III (or) Higher
* Ram: 64MB (or) Higher
* Hard disk: 80GB (or) Higher

**CHAPTER # 3**

**REQUIREMENTS AND ANALYSIS**

**Problem Definition:-**

The traditional System of newspaper delivery is very hectic, time-consuming, and labor-intensive.

* As paper backed newspapers and articles can be easily destroyed leaving the user with no information.
* Many people like to read a only specific portion of the newspaper, it cost a lot of paper wastage.
* Exam preparation Aspirants need newspapers daily for their preparation and articles.

**Requirement Specification:-**

**Number of Modules:**The WhatsApp News application has two modules in our project.

 User

Admin

**User:-**Users can receive the paper on their WhatsApp application daily at the same time.

**Admin:-**

* Admin is the super user of the application who manages everything on the application.
* Has the right to change the code
* and also add additional modules to the existing application

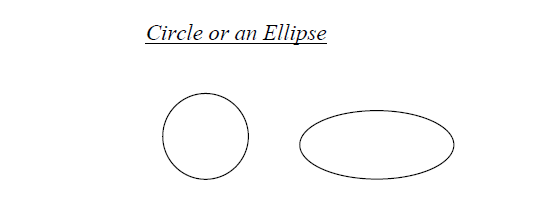
**Data Flow Diagram (DFD)**

 A data flow diagram (DFD) maps out the flow of information for any process or system.

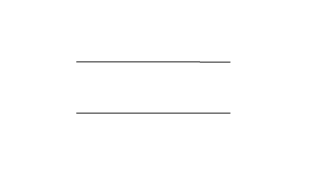
 It uses defined symbols like rectangles, circles, and arrows, plus short text labels, to show data inputs, outputs, storage points, and the routes between each destination.   
  
 Data flow diagrams are used to graphically represent the flow of data in a business information system.   
  
 DFD describes the processes that are involved in a system to transfer data from the input to the file storage and report generation.   
  
 Data flow diagrams can be divided into logical and physical.   
  
 The logical data flow diagram describes the flow of data through a system to perform certain functions of a business   
  
 The physical data flow diagram describes the implementation of the logical data flow.

**Symbols used in Data Flow Diagram are as follows:-**

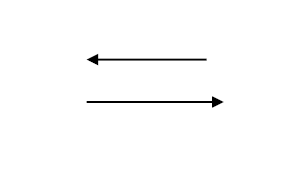
**1. Process**: - Represents process that transforms incoming data flow to outgoing data flow.



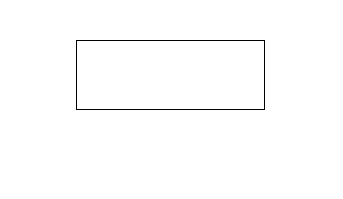
**2**.**Data Store**:- Data stores are repositories of data in the system. They are sometimes referred to as files.



**3**. **Dataflow**: - Dataflow is pipelined through which information flows.



**4**. **External entity**:-External entities are sources and destinations of the system’s inputs and outputs



**Context Free Diagram (Zero level DFD):-**

 A context diagram gives an overview and it is the highest level in a data flow diagram, containing only one process representing the entire system. It should be split into major processes which give greater detail and each major process may further split to give more detail.

 A context-free diagram represents all the external components that may interact with the system, hence displaying the entire software system as a unit. This type of chart puts the system in the middle and surrounds it with its external parts, associated entities, and surroundings. Details of the interior structure of the system are explicitly left out.

 It is made up of a context bubble, first drawn in the middle of the chart. It is usually a circle shape that represents a conceptual boundary that encloses a group of interconnected processes and activities of a project. The nitty-gritty details of the internal structure of a system are masked in a context diagram since it is strictly a high-level view of the system. This process is called information hiding.

 Context diagrams are an excellent tool for facilitating brainstorming among those who design and analyze them.

 A context diagram is suitable for noting omissions and blunders in a business plan or project requirements. Hence you can make necessary corrections and adjustments before the project execution and reduce project risks.

 It also makes a skeleton, simply and straightforwardly, of the project’s scope. It enhances the revelation and confirmation of advanced events that initiate the entire project’s processes like the external components, inputs to and outputs from the project process, and initial sub-process requirements.

 Below is the level 0 data flow diagram of our proposed system:-

**diagram**

**LEVEL 0 DATA FLOW DIAGRAM**

**LEVEL 1 DATA FLOW DIAGRAM**

**The Level 1 Data Flow Diagram is the extended version of the level 0 data flow diagram.**

Below is the level 1 data flow diagram of our proposed system:-

**LEVEL 1 DFD FOR USER**

**diagram**

**LEVEL 1 DFD FOR admin:-**

**diagram**

**Use Case Diagram:-**

A use case diagram is a dynamic or behavior diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use cases are a set of actions, services, and functions that the system needs to perform. In this context, a "system" is something being developed or operated, such as a website. The "actors" are people or entities operating under defined roles within the system.

Use case diagrams are valuable for visualizing the functional requirements of a system that will translate into design choices and development priorities.

They provide a good high-level analysis from outside the system. Use case diagrams specify how the system interacts with actors without worrying about the details of how that functionality is implemented.

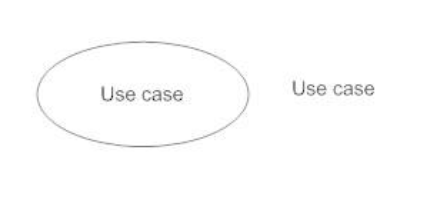
**System:-**

Draw your system's boundaries using a rectangle that contains use cases. Place actors outside the system's boundaries.



**Use Case:-**

Draw use cases using ovals. Label the ovals with verbs that represent the system's functions.



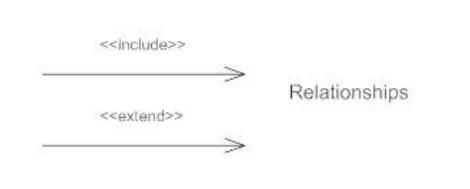
**Actors:-**

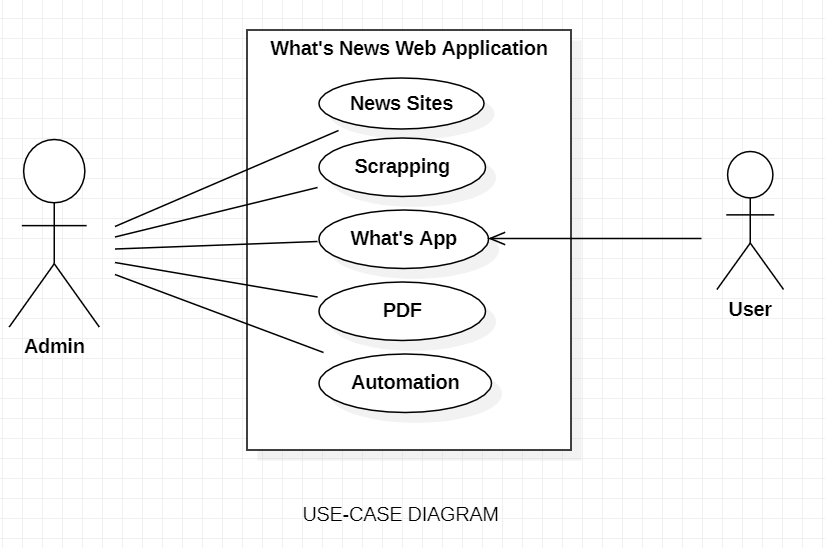
Actors are the users of a system. When one system is the actor of another system, label the actor system with the actor stereotype.



**Relationships:-**

Illustrate relationships between an actor and a use case with a simple line. For relationships among use cases, use arrows labeled either "uses" or "extends." A "uses" relationship indicates that one use case is needed by another to perform a task. An "extends" relationship indicates alternative options under a certain use case.





**Entity Relationship Diagram:-**

ER Diagram stands for Entity Relationship Diagram, also known as ERD is a diagram that displays the relationship of entity sets stored in a database. In other words, ER diagrams help to explain the logical structure of databases. ER diagrams are created based on three basic concepts: entities, attributes, and relationships.

ER Diagrams contain different symbols that use rectangles to represent entities, ovals to define attributes, and diamond shapes to represent relationships.

**ER Diagrams Symbols & Notations: -**

**Entity Relationship Diagram Symbols & Notations** mainly contain three basic symbols which are rectangle, oval, and diamond to represent relationships between elements, entities, and attributes. Some sub-elements are based on the main elements in ERD Diagram. ER Diagram is a visual representation of data that describes how data is related to each other using different ERD Symbols and Notations.

**Following are the main components and their symbols in ER Diagrams:**

 **Rectangles:** This Entity Relationship Diagram symbol represents entity types

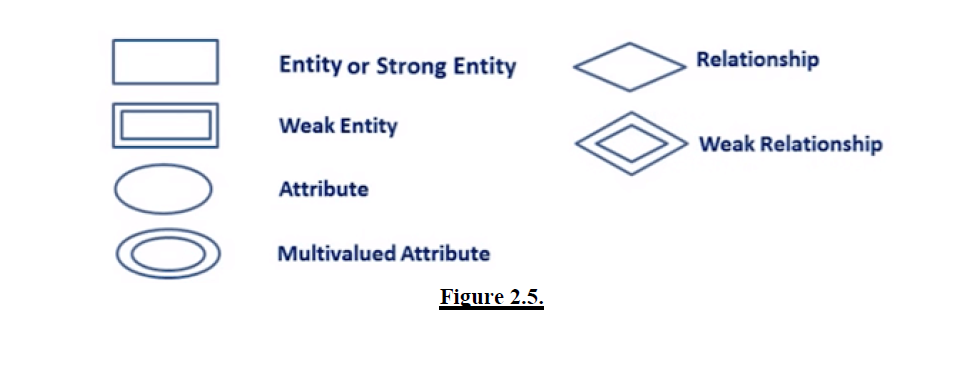
 **Ellipses:** Symbols represent attributes

 **Diamonds:** This symbol represents relationship types

 **Lines:** It links attributes to entity types and entity types with other relationship types

 **Primary key:** attributes are underlined

 **Double Ellipses:** Represent multi-valued attributes



**Components of the ER Diagram: -**

This model is based on three basic concepts:

 Entities

 Attributes

 Relationships

**er diagram**

**Research Methodology:-**

**Planning Phase:-**

During this phase, the backbones of the project are analyzed and identified to provide a clear picture of the project's feasibility and significance besides serving as the basic guideline throughout the project. The backbones refer to:

 Problems that need to be addressed throughout the project.

 Scope of the project to enable completion of the project within the duration provided.

 Main goals of the project and criteria for the successful goals.

 Potential tools for developing the solution.

**Data gathering and analyzing phase:-**

Upon project approval, data gathering and analysis are executed to support the initial planning outcome and provide opportunities to discover additional information related to the project.

**Research on the current system:-**

Details on the current WhatsApp Status are researched, analyzed, and documented in the report. Research is done via article and journal reading on WhatsApp using e-resources provided and qualitative open-ended interviews. The purpose of the action is to identify the details below:

 Differences and similarities between the current system and the proposed system.

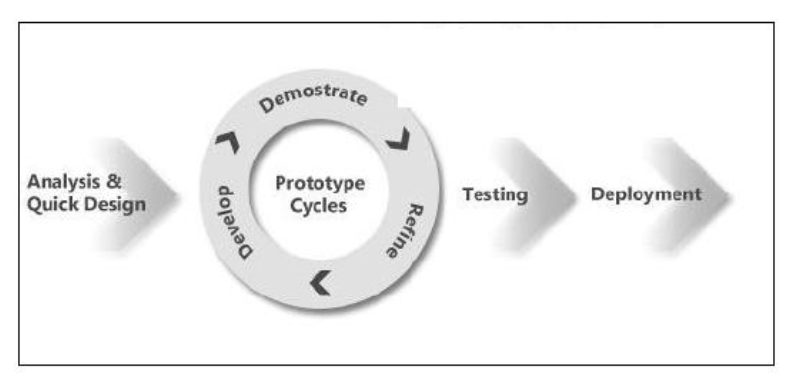
 The system architecture of the current WhatsApp can be adapted to the project.

 Potential features that can be implemented in the project such as mobility and both offline and online storage.

 Identify tools and components for developing the proposed system.

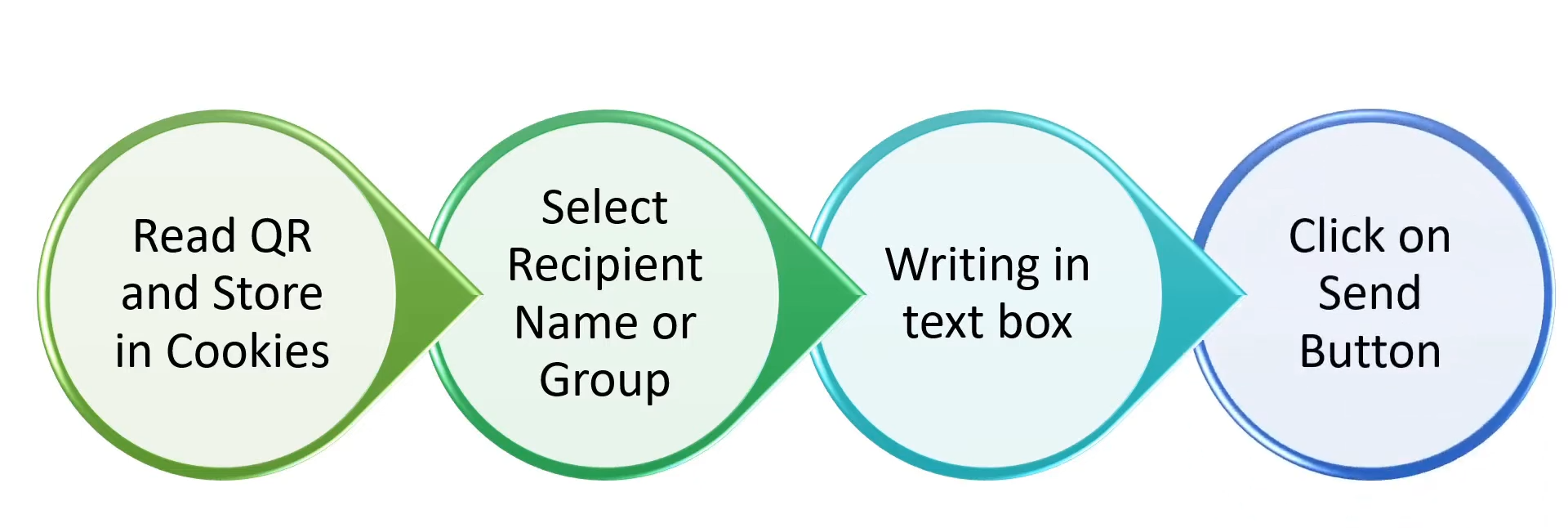
**Development Phase:-**

For this project, the main methodology used is Rapid Application Development (RAD) methodology using an incremental prototyping approach. RAD focuses more on prototyping and less on planning thus allowing faster application development and refining. The methodology is chosen after thorough research on potential methodology for this project after consideration of the project’s limitations and consultation from experts.



**System architecture design:-**

The figure shows the final system architecture for the project. The system overview will serve as the core structure for the system for future enhancement and feature implementation.



**Development Methodology:-**

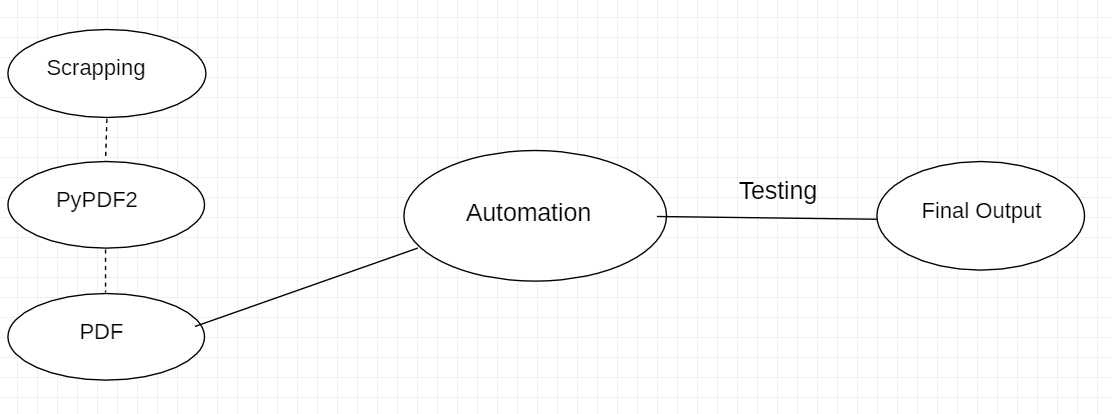
As the system adapted RAD with an incremental prototyping approach, the project is divided into three (3) main component prototypes which are:

a. Scrapping

b. PDF conversion

c. Automation

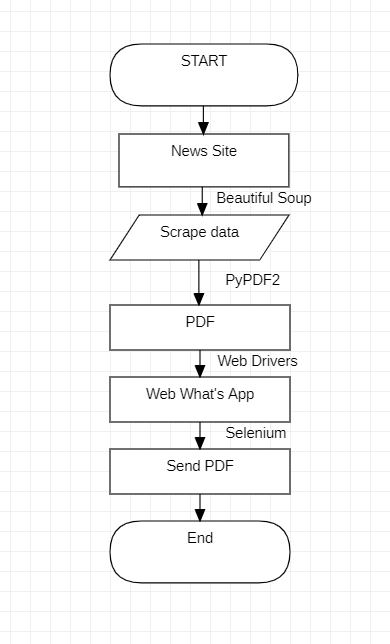
Each component is developed one at a time and undergoes individual tests to satisfy the criteria determined. At the end of a prototyping cycle, all prototypes will be merged and linked together, and tested.



**Flowchart representation of the Project: -**

This figure shows the pictorial representation of how the project follows with steps: -

diagram



**CHAPTER # 6**

**Implementation**

**Frontend:-**

**1. WhatsApp Web:-**

WhatsApp is a popular messaging application that allows users to send text messages, voice messages, make voice and video calls, and share files with friends and family over the internet. It was founded in 2009 by Jan Koum and Brian Acton and was later acquired by Facebook in 2014.

One of the reasons for its widespread popularity is its easy-to-use interface and its availability across multiple platforms, including iOS, Android, Windows, and web browsers. WhatsApp also offers end-to-end encryption, which means that messages sent between users are secured and can only be read by the sender and receiver.

WhatsApp has evolved to become a popular tool for businesses to communicate with customers, as well as a platform for sharing news and information with large groups of people. It has over 2 billion active users worldwide, making it one of the most popular messaging apps in the world.



**Back end: -**

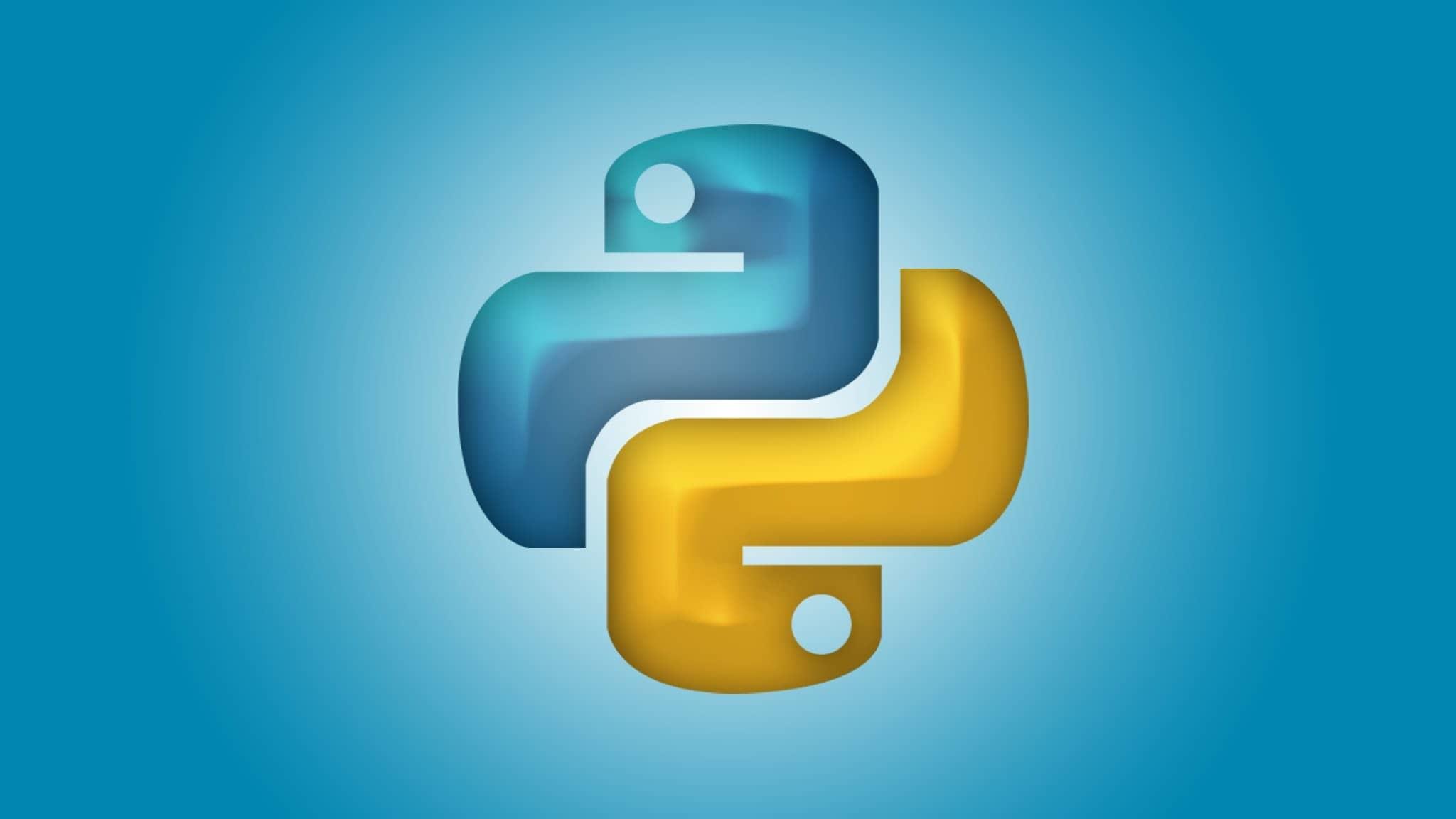
**Python:-**

Python is a high-level, interpreted programming language that was first released in 1991 by Guido van Rossum. It is designed to be easy to read, write, and maintain, making it a popular choice for beginners and experienced programmers alike.

Python has a wide range of applications, including web development, data analysis, scientific computing, artificial intelligence, machine learning, and automation. It has a large and active community of developers who contribute to its libraries, frameworks, and tools, making it a powerful and versatile language.

One of the key features of Python is its simplicity, with a syntax that emphasizes readability and reduces the cost of program maintenance. It also has a large standard library that provides many useful modules and functions for common tasks, making it easy to write complex programs quickly.

Python is an open-source language, which means that it is free to use, modify, and distribute. It is available for Windows, macOS, Linux, and other platforms, making it accessible to a wide range of users.



**Selenium:-**

Selenium is a popular open-source automation tool for testing web applications. It provides a suite of tools and libraries for automating web browsers, which can be used to simulate user actions, perform automated testing, and perform web scraping.

Selenium supports several programming languages, including Python, Java, C#, and Ruby, among others, making it a versatile automation tool. It allows testers and developers to write automated test scripts that can simulate user interactions with web applications, such as clicking buttons, filling out forms, and navigating through pages.

One of the key benefits of using Selenium for automation is that it supports cross-browser testing, which means that tests can be run on multiple browsers, such as Chrome, Firefox, Safari, and Internet Explorer. This helps ensure that web applications function correctly across different platforms and browsers.

Selenium also supports integration with other testing tools and frameworks, such as TestNG, JUnit, and NUnit, making it easy to incorporate automated testing into a larger software development process.

Overall, Selenium automation can help improve the efficiency and reliability of testing web applications, leading to faster and more accurate software development.



**PyPdf:-**

PyPDF is a Python library that provides a set of tools for working with PDF files. It allows users to extract information from PDFs, manipulate PDF pages, and merge multiple PDFs into a single document.

PyPDF is an open-source library that can be installed using pip, the Python package manager. Once installed, it can be imported into a Python script and used to perform a wide range of PDF-related tasks.

Some of the key features of PyPDF include:

1. PDF parsing: PyPDF can be used to extract text, images, and other data from PDFs. It can also be used to search for specific text or patterns within a PDF.

2. PDF manipulation: PyPDF can be used to rotate, crop, and resize PDF pages. It can also be used to add watermarks, bookmarks, and annotations to PDFs.

3. PDF merging: PyPDF can be used to merge multiple PDFs into a single document. This is useful for combining reports, invoices, and other PDF-based documents.

4. PDF encryption: PyPDF can be used to encrypt PDFs with passwords, preventing unauthorized access to sensitive information.

Overall, PyPDF is a powerful and flexible tool for working with PDFs in Python and is particularly useful for data extraction and manipulation tasks.

**Testing:-**

System testing is a series of different tests whose primary purpose is to fully exercise computer-based systems.

We can say that it will run according to its specifications and in the way users expect. Special test data are input for processing, and the results are examined. A limited number of users may be allowed to use the system so that analysts can see whether they try to use it in unforeseen ways. It is desirable to discover any surprises before the organization implements the system and depends on it.

▪We follow Black Box testing

▪ Black box testing attempts to find errors in the following

▪ Incorrect or missing function

▪ Interface errors

▪ Errors in data structure

▪ Initialization and termination errors

**CHAPTER # 7**

**Coding**

**1. Scrapping Code:-**

def extract\_month(mn):

    if(mn==1):

        mon = "jan"

    elif(mn==2):

        mon = "feb"

    elif(mn==3):

        mon = "mar"

    elif(mn==4):

        mon = "apr"

    elif(mn==5):

        mon = "may"

    elif(mn==6):

        mon = "jun"

    elif(mn==7):

        mon = "jul"

    elif(mn==8):

        mon = "aug"

    elif(mn==9):

        mon = "sep"

    elif(mn==10):

        mon = "oct"

    elif(mn==11):

        mon = "nov"

    elif(mn==12):

        mon = "dec"

    return mon

def scrape():

    import time

    from PyPDF2 import PdfFileMerger, PdfFileReader

    import os

    current\_time = time.localtime(time.time())

    mon = extract\_month(current\_time.tm\_mon)

    year = current\_time.tm\_year

    date = current\_time.tm\_mday

    from bs4 import BeautifulSoup as BS

    import requests

    page = requests.get("https://epaper.dailyexcelsior.com/")

    pageCount=15

    if(page.status\_code==200):

        soup = BS(page.content, 'html.parser')

        s = soup.find('div', id="scroller")

        pageCount=len(s.select('ul', class\_='epapertest')[0].select('li'))

    for i in range(1, pageCount):

        url = '''https://epaper.dailyexcelsior.com/epaperpdf/{}/{}/{}{}{}/page{}.pdf'''.format(year, mon, str(year)[:2], mon, str(date).zfill(2), i)

        r = requests.get(url)

        f = open("News/temp/{}.pdf".format(i), "wb")

        f.write(r.content)

        f.close()

        time.sleep(2)

    pdfs = set()

    for element in os.listdir("News/temp/"):

        pdfs.add(int(element.split(".")[0]))

    pdfs = [str(i)+".pdf" for i in list(pdfs)]

    merger = PdfFileMerger()

    for pdf in pdfs:

        merger.append("News/temp/" + pdf)

    merger.write("News/temp/dailyexcelsior.pdf")

    merger.close()

**2. Automation Code:-**

def send\_message(message, pdf\_loc, contact\_list):

    from selenium import webdriver

    from selenium.webdriver.common.by import By

    from selenium.webdriver.support.ui import WebDriverWait

    from selenium.webdriver.support import expected\_conditions as EC

    from selenium.webdriver.common.keys import Keys

    import time

    import sys, os

    from .config import CHROME\_PROFILE\_PATH

    options = webdriver.ChromeOptions()

    options.add\_argument(CHROME\_PROFILE\_PATH)

    #For Local or EC2

    driver = webdriver.Chrome(options=options)

    driver.get("https://web.whatsapp.com")

    try:

        for contact in contact\_list:

            search\_xpath = '//div[@contenteditable="true"][@data-tab="3"]'

            search\_box = WebDriverWait(driver, 500).until(

                EC.presence\_of\_element\_located((By.XPATH, search\_xpath))

            )

            search\_box.clear()

            search\_box.send\_keys(contact)

            contact\_xpath = f'//span[@title="{contact}"]'

            contact\_title = WebDriverWait(driver, 500).until(

                EC.presence\_of\_element\_located((By.XPATH, contact\_xpath))

            )

            contact\_title.click()

            # Send Text

            input\_xpath = '//div[@contenteditable="true"][@data-tab="6"]'

            input\_box = WebDriverWait(driver, 50).until(

                EC.presence\_of\_element\_located((By.XPATH, input\_xpath))

            )

            input\_box.send\_keys(message)

            input\_box.send\_keys(Keys.ENTER)

            # Send Attachment

            attachment\_xpath = '//span[@data-testid="clip"][@data-icon="clip"]'

            attachment\_box = WebDriverWait(driver, 500).until(

                EC.presence\_of\_element\_located((By.XPATH, attachment\_xpath))

            )

            attachment\_box.click()

            document\_box = driver.find\_element\_by\_xpath('//input[@accept="\*"]')

            abs\_pdf\_loc = os.path.abspath('..') + "/" + pdf\_loc + "dailyexcelsior.pdf"

            document\_box.send\_keys(abs\_pdf\_loc)

            time.sleep(3)

            send\_btn = WebDriverWait(driver, 500).until(EC.presence\_of\_element\_located((By.XPATH, '//span[@data-icon="send"]')))

            send\_btn.click()

            time.sleep(10)

    finally:

        driver.quit()

**3. Deployment Code:-**

def job(scheduled\_time):

    import os

    import shutil

    from News.news import scrape

    from Whatsapp. whatsapp import send\_message

    os.mkdir("News/temp")

    scrape()

    contact\_list = ["Project news"]

    message = f"This message is autogenerated and is scheduled at {scheduled\_time}. You will daily receive this message at the same time. Happy reading!"

    pdf\_loc = "Whatsapp News/News/temp/"

    send\_message(message, pdf\_loc, contact\_list)

    shutil.rmtree("News/temp/", ignore\_errors = False)

def schedule\_job():

    import schedule

    import time

    scheduled\_time = "14:19"

    schedule.every().day.at(scheduled\_time).minutes.do(job, scheduled\_time)

    while True:

        schedule.run\_pending()

        time.sleep(10)

if \_\_name\_\_ == "\_\_main\_\_":

    schedule\_job()

**CHAPTER # 7**

**Testing**

Software Testing is the process used to help identify the correctness, completeness, security, and quality of developed computer software. Testing is a process of technical investigation, performed on behalf of stakeholders, that is intended to reveal quality-related information about the product concerning the context in which it is intended to operate. This includes but is not limited to, the process of executing a program or application with the intent of finding errors. Quality is not absolute; it is valuable to some people. With that in mind, testing can never completely establish the correctness of arbitrary computer software; testing furnishes a criticism or comparison that compares the state and behavior of the product against a specification. The important point is that software testing should be distinguished from the separate discipline of Software Quality Assurance (SQA), which encompasses all business process areas, not just testing.

There are many approaches to software testing, but effective testing of complex products is essentially a process of investigation, not merely a matter of creating and following routine procedures. One definition of testing is "the process of questioning a product to evaluate it", where the "questions" are operations the tester attempts to execute with the product, and the product answers with its behavior in reaction to the probing of the tester[citation needed]. Although most of the intellectual processes of testing are nearly identical to that of review or inspection, the word testing is connoted to mean the dynamic analysis of the product—putting the product through its paces. Some of the common quality attributes include capability, reliability, efficiency, portability, maintainability, compatibility, and usability. A good test is sometimes described as one which reveals an error; however, more recent thinking suggests that a good test reveals information of interest to someone who matters within the project community.

**Introduction:-**

Software testing can be a sub-field of Software Quality Assurance but typically exists independently (and there may be no SQA areas in some companies). In SQA, software process specialists and auditors take a broader view of software and its development. They examine and change the software engineering process itself to reduce the number of faults that end up in the code or deliver faster.

Regardless of the methods used or the level of formality involved the desired result of testing is a level of confidence in the software so that the organization is confident that the software has an acceptable defect rate. What constitutes an acceptable defect rate depends on the nature of the software. An arcade video game designed to simulate flying an airplane would presumably have a much higher tolerance for defects than software used to control an actual airliner.

A problem with software testing is that the number of defects in a software product can be very large, and the number of configurations of the product is larger still. Bugs that occur infrequently are difficult to find in testing. A rule of thumb is that a system that is expected to function without faults for a certain length of time must have already been tested for at least that length of time. This has severe consequences for projects to write long-lived reliable software. Unit tests are maintained along with the rest of the software source code and generally integrated into the build process (with inherently interactive tests being relegated to a partially manual build acceptance process). The software, tools, samples of data input and output, and configurations are all referred to collectively as a test harness.

**Test Case Documentation:-**

**Stlc:-**

 Test Planning.

 Test Development.

 Test Execution.

 Result Analysis.

 Bug-Tracing.

 Reporting.

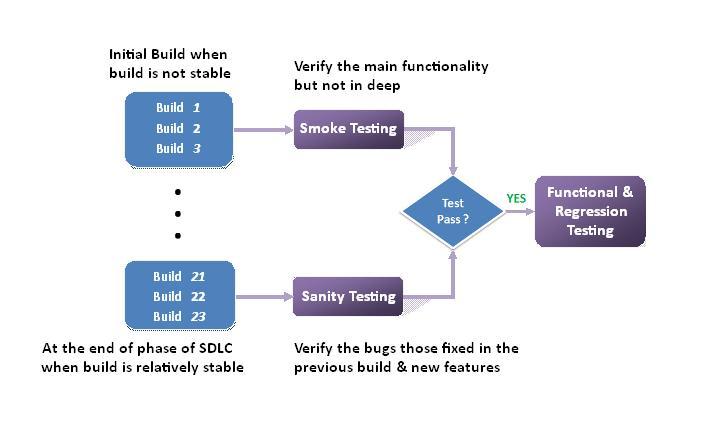
 Microsoft windows-standards

Testing follows a few steps and also has a different level of testing with different types

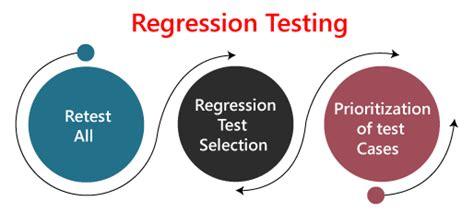
**Types Of Testing:-**

 **Smoke Testing**: this is the process of initial testing in which the tester looks for the availability of all the functionality of the application to perform detailed testing on them. (The main check is for available forms)

 **Sanity Testing: this** is a type of testing that is conducted on an application initially to check for the proper behavior of an application that is to check all the functionality that is available before the detailed testing is conducted on them.



 **Regression Testing:** is one of the best and most important testing. Regression testing is the process in which the functionality, which is already tested before, is once again tested whenever some new change is added to check whether the existing functionality remains the same.

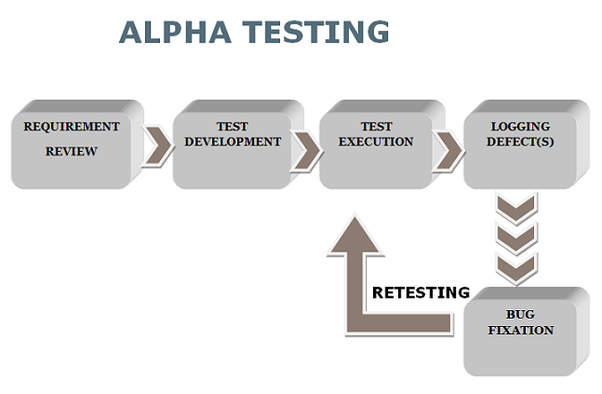


 **Re-Testing: this** is the process in which testing is performed on some functionality which is already been tested before to make sure that the defects are reproducible and to rule out environmental issues if at all any defects are there.

 **Static Testing:** is the testing, which is performed on an application when it is not been executed. ex: GUI, Document Testing

 **Dynamic Testing:** is the testing that is performed on an application when it is being executed. ex: Functional testing.

 **Alpha Testing:** it is a type of user acceptance testing, which is conducted on an application when it is just before released to the customer.



 **Beta-Testing:** it is a type of UAT that is conducted on an application when it is released to the customer when deployed into the real-time environment and being accessed by the real-time users.

 **Monkey Testing:** is the process in which abnormal operations, beyond capacity operations, are done on the application to check its stability of it despite the user’s abnormal behavior.

 **Compatibility testing:** it is the testing process in which usually the products are tested on environments with different combinations of databases (application servers, browsers…etc.) To check how far the product is compatible with all these environments' platform combinations.

 **Installation Testing:** it is the process of testing in which the tester tries to install or try to deploy the module into the corresponding environment by following the guidelines produced in the deployment document and checking whether the installation is successful or not.

 **Adhoc Testing:** Adhoc Testing is the process of testing which unlike formal testing where in test case document is used, without that test case document testing can be done for an application, to cover that testing of the future which is not covered in that test case document. Also, it is intended to perform GUI testing which may involve cosmetic issues.

**Testing Methodology:-**

 Black box Testing:

 White box Testing.

 Gray Box Testing

.

**Testing Levels:-**

 Unit Testing.

 Module Testing.

 Integration Testing.

 System Testing.

 User Acceptance Testing

**CHAPTER # 7**

**Result**

result output

**ADVANTAGES, APPLICATIONS, AND LIMITATIONS**

**ADDITIONAL MODULE**

**FAKE NEWS DETECTION USING MACHINE LEARNING**

**Abstract**

Counterfeit news is the purposeful spread of deception or promulgation by conventional news media and through online media. Such news stories can come in numerous structures, including accidental mistakes submitted by news aggregators, inside and out bogus stories, or tales that are created to deceive and impact per user's assessment. Likewise, since bogus data can spread so quickly, not just it can hurt individuals yet additionally can be hindering to tremendous enterprises and financial exchanges.

In the previous decade, we have seen a dramatic development of data accessible on the web. It is getting difficult to interpret valid from the bogus. We need to know how much of what we read on supposedly credible news sites is trustworthy. False information can cause panic among citizens. Likewise lies can be utilized to control different people's choices for casting a ballot bid or basically whatever else that can have enduring consequences. Bogus data spreads extremely quickly, this is shown by the way that when one phony news site is brought down another immediately has its spot. Besides, it is fit for demolishing the equilibrium of the news environment. Political plans and control are one of the numerous intentions since counterfeit news is created.

Today, fortunately, we have advances in machine learning and Language Processing (NLP) instruments that offer incredible guarantees for specialists to construct frameworks that could naturally identify counterfeit news. Hence it would be beneficial to discuss the methods of detecting online deceptions. Our undertaking can be essentially utilized by any media organization to foresee if the coursing news is phony. The interaction should be possible consequently without having people physically audit a huge number of information-related articles.

Approaching the problem from a purely NLP perspective, it will be possible for us to predict whether the news is fake or real based only on its content. A significant piece of the objective is to think about and report the outcomes from numerous distinctive model executions, and present an investigation of the discoveries.

We will use several architectures like naïve based classifiers and random forests to come up with an accurate prediction.

**LIST OF TABLES**

**Sl. No. TABLE NAME**

1 Project Schedule

2 Unit Testing

3 Risk Analysis

4 Verification and Validation

5 Integration testing

6 Mc Call’s Quality factors

**LIST OF FIGURES**

**Figure. No. FIGURE NAME**

1 Use-Case Diagram

2 State/Activity Diagram

3 Deployment Diagram

4 ER-model

5 Relational model

6 Sequence Diagram

7 Architecture Diagram

**CHAPTER # 1**

**Introduction**

In various venues, fake news is causing a variety of concerns, ranging from humorous articles to falsified news and planned government propaganda. A significant piece of the objective is to think about and report the outcomes from numerous distinctive model executions, and present an investigation of the discoveries.

The relevance of disinformation in American political discourse has received a lot of attention recently, especially in the aftermath of the presidential election in the United States. The phrase 'fake news came to be used to designate factually erroneous and deceptive items that were published primarily for the goal of generating revenue through page views. This paper appears to have created a model that can reliably estimate the chances of a given article being false news.

After media attention, Facebook was at the center of much criticism. They've already included a tool that flags counterfeit news on a website when a client comes across such a page, and the company has expressed openly that they're chipping away at a programmed system to distinguish false stories, it is a troublesome undertaking. Since counterfeit news shows up on the two finishes of the political range, a given calculation should be politically unprejudiced while as yet giving real news sources on the two closures of the range equivalent weight. Furthermore, it is a difficult question of legitimacy. Be that as it may, it is important to comprehend what counterfeit news is and what different methodologies have tackled this problem. It is important to concentrate on how AI and the investigation of characteristic dialects permit us to identify bogus news.

There is a Kaggle contest called the "Bogus News Challenge," and Facebook is utilizing artificial intelligence to sift counterfeit reports through clients' channels. Battling counterfeit news is a conventional book classification project with a basic proposition. Whether it is feasible to make a model that can recognize "genuine" and "counterfeit" news or not? Thus, a proposed exertion on accumulating a dataset of both phony and authentic news and utilizing a random forest classifier to foster a model to sort an article as bogus or genuine dependent on its words and expressions.

The fundamental objective is to identify the bogus news, which is an old-style text rating issue with a straightforward proposition. It is important to foster a model for recognizing "genuine" and "counterfeit" news.

**CHAPTER # 2**

**Literature Survey**

**2.1 Literature Survey:-**

**Title 1: Automatic deception detection: Methods for finding fake news**

Authors: Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November).

Abstract: The main purpose of their study was to construct linguistic cue techniques, word approach, rhetorical structure and discourse analysis, network analysis methods, and SVM classification. SVM classification techniques were provided. These are text-based models, which offer very little or no improvement on existing approaches.

Result: Random forest requires numerous features and the survey showed the highest accuracy achievement of 95%.

**Title 2: Weakly supervised learning for fake news detection on Twitter**

Authors: Helm Stetter, S., & Paulheim, H. (2018, August).

Abstract: Fake news identification on Twitter has been poorly controlled, and every Tweet/Post has been categorized as a binary classification issue. The grades are based exclusively on the post/tweet source. Authors utilize Twitter API, and DMOZ, and utilized techniques such as naïve bays, SVM, XG boost, and neural nets. The author uses data sets manually. The statistics reveal 15% fraudulent tweets, 45% actual tweets, and remaining unresolved postings.

Result: Results suggest that the rest of the post was undecided: 15 percent bogus tweets, 45 percent real tweets.

**Title 3: Automatic Online Fake News Detection Combining Content and Social Signals.**

Authors: Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May).

Abstract: Facebook Messenger's chatbot implemented it. Three various datasets of Facebook Italian news posts were used. The Boolean crowd-sourcing algorithms were implementable, both for content-based approaches with social and content signals.

Result: Three various datasets of Facebook Italian news posts were used. The Boolish crowdsourcing algorithms are used for both content-based approaches using social and content signals.

**2.2 Problem Description:-**

Phishing is tough to trace or defend against since it does not seem malicious at first look. Also, the cost and difficulty of deploying such attacks are decreasing substantially. As it is challenging to know if an URL is safe or not, we created a model which serves in classifying URLs into safe or legitimate classes.

URLs contain components of its page and hence are used to identify the purpose of the web page without checking its actual content. We suggest a framework based on a deep learning algorithm (Bi-Directional Long-Term Memory) that predicts the state of the URL without the use of domain knowledge or manual extraction with more important accuracy points than already existing systems.

**2.3 Requirements Gathering:-**

Prototyping: Specifications for prototypes consist of preparing and then creating a model that can be provided for consumers to "play," check and change.

Interactions: Interviews are the primary means of gathering information where the program manager can meet face-to-face with specific clients or subject matter specialists Interviews are the primary means of gathering information where the program is located.

Focus Groups: A synergistic conversation is taking place between individuals who are representations of consumers or buyers about the requirements of the company.

**2.4 Requirements Analysis:-**

**2.4.1 Functional Requirements**

Collecting Dataset: Creating a dataset for the training of the models is the primary task. Datasets can be created manually and then augmented.

Pre-Processing: By sending the dataset to the model, the dataset should be pre-processed. This includes splitting the data into three folds.

Training Models: To do an optimal phishing detection model, they should be trained and validated.

Detecting Fake News: Training of models to detect real news from the pool of fake news.

**2.4.1 Non-Functional Requirements**

• Accuracy

• Less training time

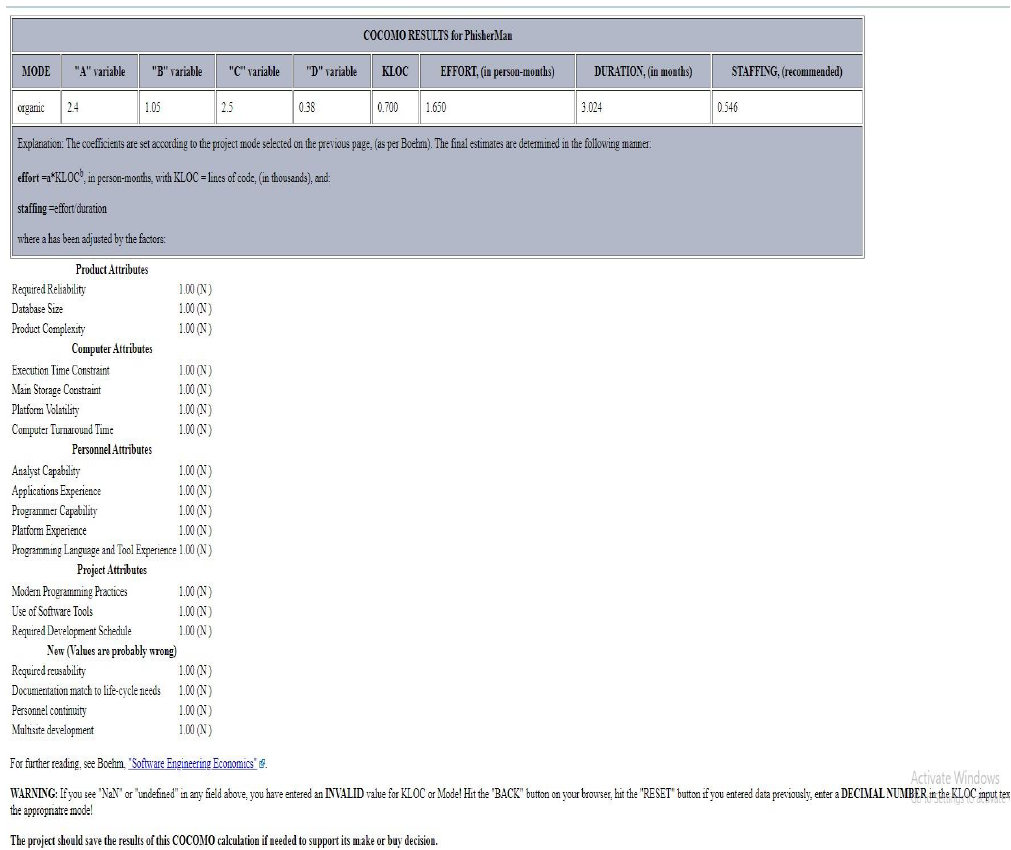
• Simple architecture

**2.2 Data Source:-**

There are two sorts of articles in the dataset: false News and actual News. This data set has been collected from sources in the real world and the real articles have been gathered from Reuters.com (News website). The false news pieces have been collected from various sources. The bogus news stories were collected on websites flagged by Political Facts (a U.S. organization of fact-checking) and Wikipedia.

The dataset is made up of two CSV files. The first file, "True.csv," contains almost 12,600 items from reuters.com. The second file, “Phony.csv,” comprises almost 12,600 items from various fake news outlet sites. Every article contains the title of the article, text, type, and date of publication of the article.

**2.3 Cost Estimation:-**

*Figure 3 COCOMO MODEL*

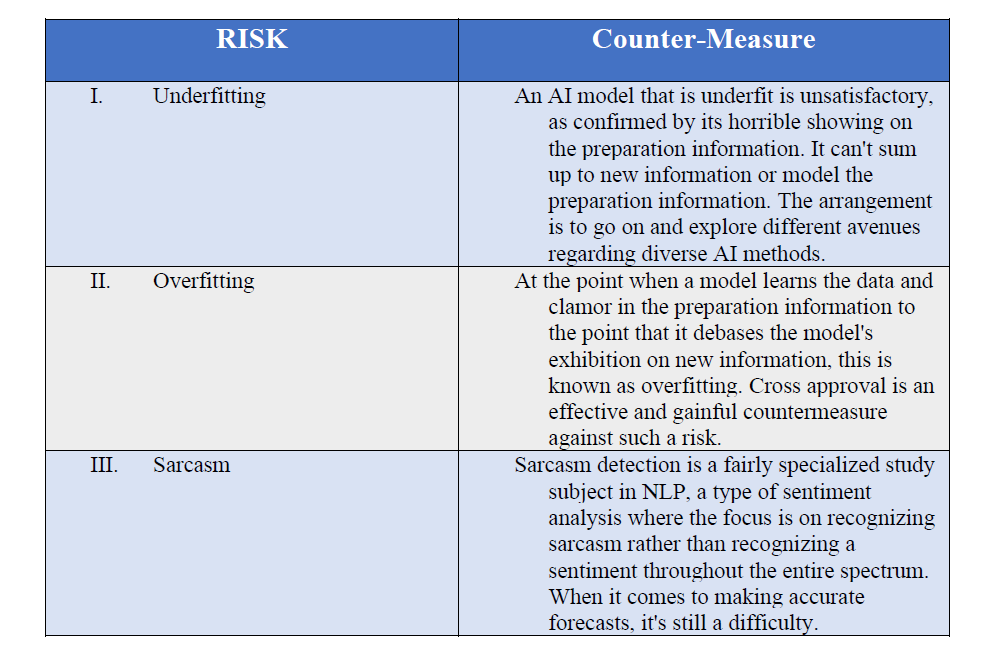
The Cocomo model is applied in our project, as our software

project is organic Effort Applied (E) *= a \** ()#

Development Time (D) *= c \** ()%

**2.4 Risk Analysis:-**

For our project, we have the following three major risks along with their countermeasures exercised.



Along with this, there is always the risk of other factors such as:

* User input is invalid
* Quality Issues
* Hardware Issues

**CHAPTER # 3**

**Software Requirement Specification**

**1. Problem Description**

The amount of information available on the internet has increased at an exponential rate. It's becoming increasingly difficult to tell what's true and what's not. We must understand how much of what we read on ostensibly reliable news sites can be trusted. False information can lead folks to panic. Also, lying may be used to influence other people's voting decisions or virtually anything else with long-term consequences. False information travels quickly, as evidenced by the fact that when one fake news site goes offline, another pops up to take its place. Fake news has the potential to destabilize the news ecosystem. One of the numerous examples is political agendas and manipulation.

**1.1.Introduction**

**1.2. Aim**

The main goal is to detect bogus news, which is a standard text classification issue with a simple proposal. It is necessary to develop a model that can distinguish between "real" and "fake" news.

**1.1.2. Document convention**

This document uses Arial as the font theme for normal content and Times as the font theme for sub-headings. The font size of the content is 11 while subheadings have a font size of 14 and they are in bold. In addition, important text in the content is highlighted by bolding the text. Furthermore, the goals for higher-level specifications are believed to be inherited from the comprehensive specifications.

**1.1.3. Product Scope**

This product is intended to detect fake news vs real news using machine learning. This product will help the user to distinguish between legitimate news and real news. This will help the users to protect themselves from malicious sites and keep the system safe. This product can predict this by using random forests, naïve based classifiers, and decision trees. Its scope is more towards detecting authentic news and preventing the spread of misinformation.

**2. Description**

This product helps to detect if the news is legitimate or not. Its main perspectives are

* + To design a machine learning model to detect phishing fakes.
  + To provide higher accuracy points as compared to other models.
  + To attempt to maintain the training time while not increasing the computation cost.

1. **Product Functions**
   * Collecting Dataset: Creating a dataset for the training of the models is the primary task. Datasets can be created manually and then augmented.
   * Pre-Processing: By sending the dataset to the model, the dataset should be pre-processed. This includes splitting the data into three folds.
   * Training Models: To do an optimal phishing detection model, they should be trained and validated.
   * Detecting phishing URLs: Training of models to detect phishing URLs from the pool of URLs consisting of legitimate and phishing URLs

**4. Characteristics**

This software can be used by anyone who wishes to be vigilant of the news articles that they are reading and find out whether they are real or not

**5. Operating environment**

This software would be available on PCs. Also, there would be a version available for upcoming smart TVs. For mobiles, it would be available on all platforms including Android, windows, ios, and Blackberry as well. It would be available on mobile as well.

**6. Design and implementation constraints**

One of the biggest constraints would be that there would be the constant requirement for internet connection like Wi-Fi networks or LAN connections to operate. As the product would be using a big database to analyze issues, it might get slow and it might get tough to get quick responses. Reverse engineering of the product is not possible unless permitted by the company.

**7. User Documentation**

There is an online tutorial available on the official website of the company. There would be a user guide as well along with the product on instructions on how to install it on various platforms.

**8. Assumptions and dependencies**

One of the assumptions is that operating devices would be having at least a minimum of 3GB RAM to handle the load of software processing. Below that, the software might crash and can also stop responding. But, with newer upcoming editions, developers would be working on making the software lighter.

**9. External Requirements**

The following documents describe the concept of computing resource requirements and standards that need to be implemented on a device to ensure the optimum operation of the program.

**Software specifications for the project:-**

Operating system: Windows 10

Coding language: Python

**10. Communication Interfaces**

Since the different components of the system are responsible for different functions and as they are dependent on each other as well, then communication between them is important. All components would be looking for each other for their data to be processed in some way. So, for this, the underlying operating system would be responsible and will carry out all the communication or data transfer processes.

**11. System Features**

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

**12. Other Non-functional Requirements**

* **Performance Requirements**

This software will require a good internet connection to connect with servers and a good processing system to give the best performance. However, developers are trying hard to make applications lighter and easier to use.

* **Safety Requirements**

There is a high chance of software crashes if the minimum requirement of RAM is not met. This can cause the device to shut down abruptly and it can even start to hang as well. So, it is recommended that users must have a minimum of 3 GB Ram to make the most of the application. This software should be bought only from sources.

* **Security Requirements**

The user must be registered with the software and software license. The software should be genuinely bought from the market or the website and the system should be registered properly. Aspirated software would not be able to access our databases and they tend to mine data on the user's device and the company would not be responsible for this.

**13. Software Quality Attributes**

* **Adaptability** – The product is very adaptable. It can work on maximum platforms, its features might not change with the change of platform.
* **Availability** – This product will be available on the official website for free.
* **Correctness** – This software uses machine learning models to predict whether the news is fake or not with an accuracy of 97%. Therefore there is correctness

**14. Business Rules**

This product will be a freely available extension on Google, after downloading the extension from google chrome and adding it to their browser the system will automatically warn the users whenever they go into a malicious website. So, we have a large target audience.

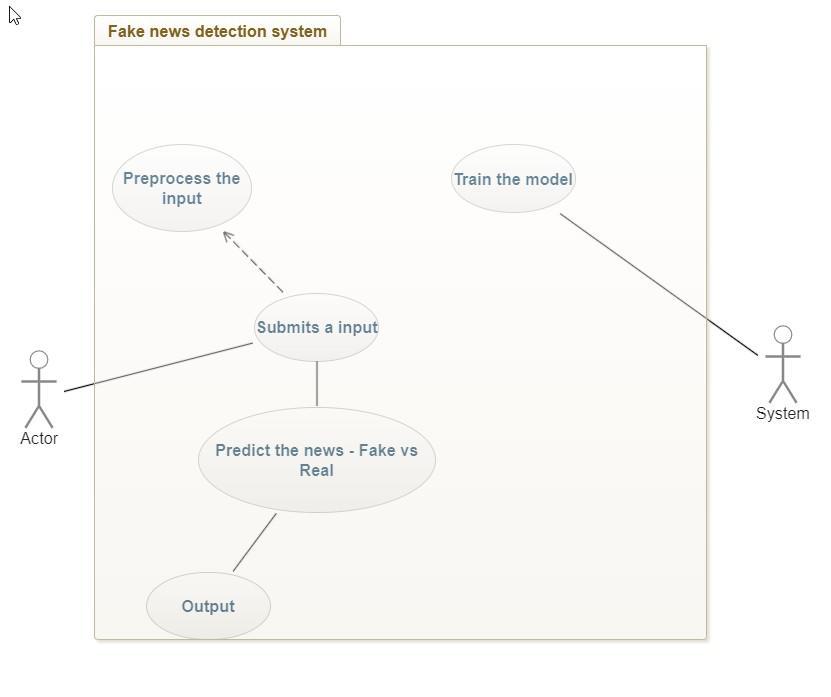
**15. Other Requirements**

In the initial stages, software is believed to take a lot of computational time so it is recommended for users to have high RAM to get the most out of the product.

1. **Use-case Diagram**

**Scenario**: User Interaction

**Actors**: An actor is an individual, entity, or external structure that plays a role with our structure in one or more encounters (actors are usually drawn as UML Stick Figures Use case diagrams).



**17. Use-case documentation**

The following four tables show the use-case documentation of all possible four scenarios.

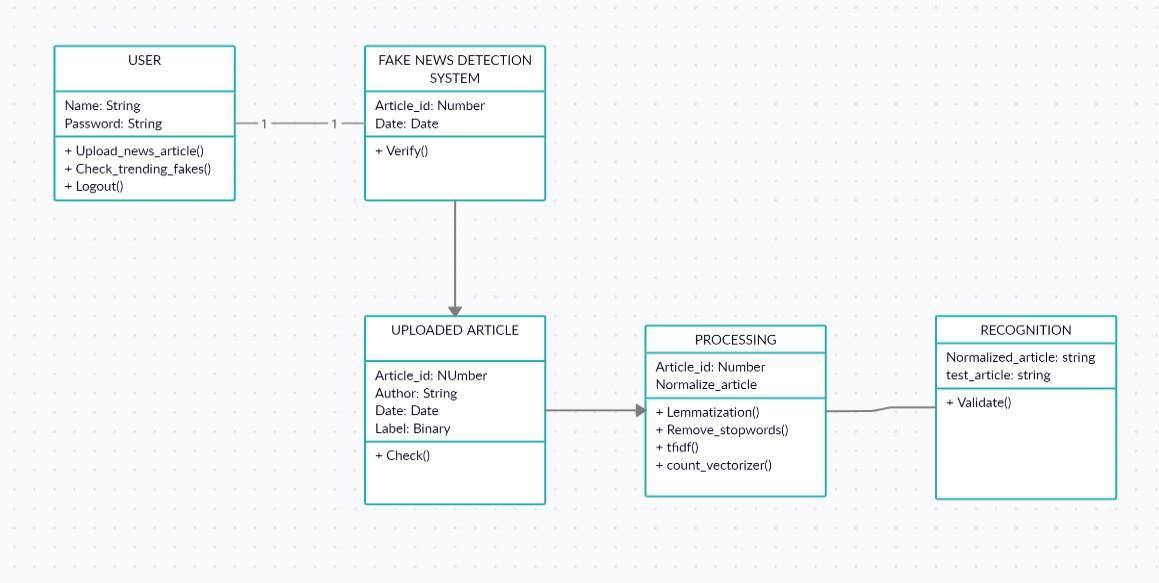
| **Use Case Name** | Accept the news article from the user along with author name, date of publication |
| --- | --- |
| **Actor** | User |
| **Pre-Condition** | Train the model |
| **Post- Condition** | None |
| **Includes** | None |
| **Extends** | None |
| **Frequency of use** | The frequency of usage can be any |
| **Normal Course of Events** | The input is entered into the screen |
| **Notes** |  |

| **Use Case Name** | Submit the input articles |
| --- | --- |
| **Actor** | user |
| **Pre-Condition** | The input of the article by the user |
| **Post- Condition** | User inputs are submitted to the system |
| **Includes** | Pre-process the input |
| **Extends** | None |
| **Frequency of use** | Can be any |
| **Normal Course of Events** | Take the input |
| **Notes** | - |

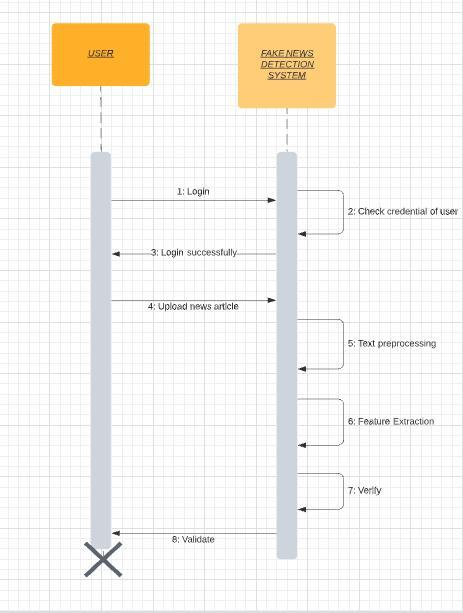
| **Use Case Name** | Predict |
| --- | --- |
| **Actor** | System, User |
| **Pre-Condition** | News articles accepted by the system |
| **Post- Condition** | The user’s input is submitted to the server |
| **Includes** | None |
| **Extends** | None |
| **Frequency of use** |  |
| **Normal Course of Events** | Takes the user input and predicts whether the news is fake or real |
| **Notes** | - |

| **Use Case Name** | Output |
| --- | --- |
| **Actor** | System, User |
| **Pre-Condition** | News articles accepted by the system |
| **Post- Condition** | Result Displayed |
| **Includes** | None |
| **Extends** | None |
| **Frequency of use** |  |
| **Normal Course of Events** | | Takes input from the news article and displays the result. | | --- | |
| **Notes** |  |

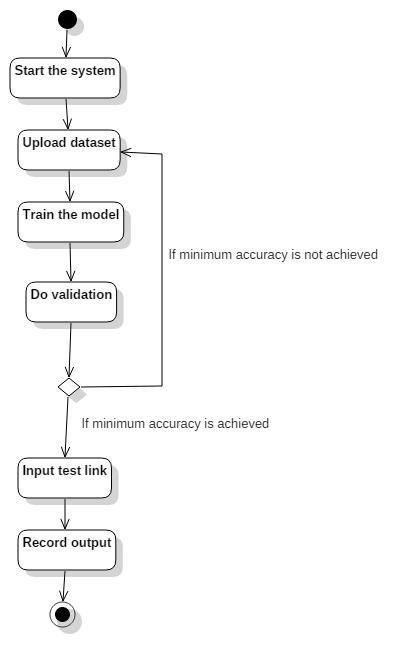
**18. Class Diagram**



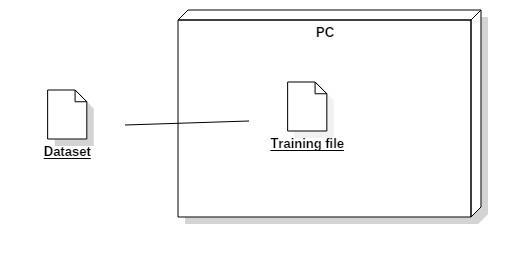
**19. Sequence Diagram**



**20. Activity/State Diagram:**



**21. Component/Deployment Diagram:**

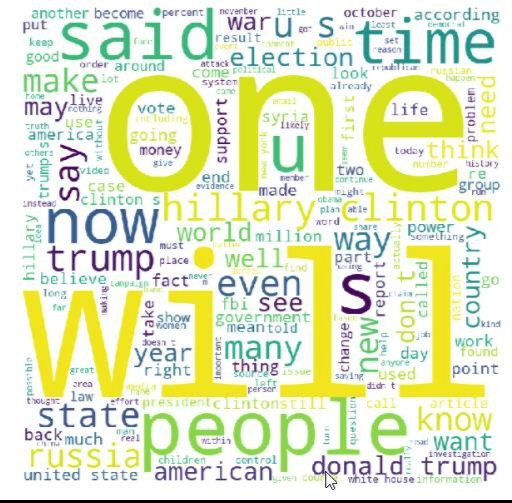


**Proof of concept:-**

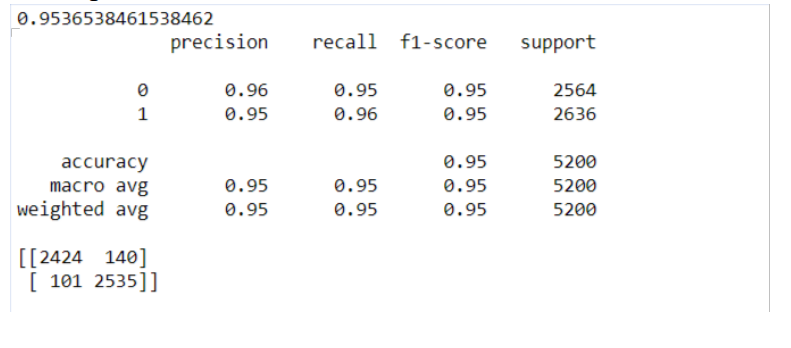
**Features of the data set:-**



**Building a word cloud:-**



**Training the model:-**



**22. Conclusion**

The model provides proof that the Support vector classifier provides higher accuracy points as compared

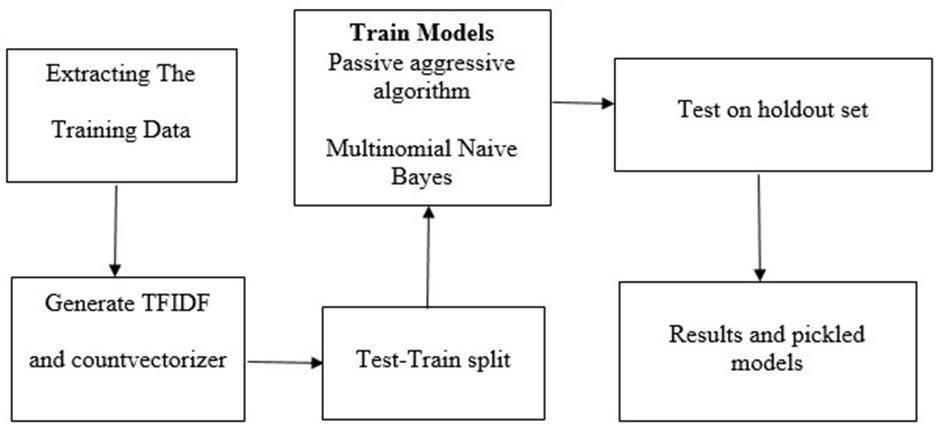
to Random forest, Naive Bayes, and Logistic Regression. We get this result as our model can change the C value, thereby identifying better-fit patterns with about the same computational capacity. With these findings,

we can safely infer that Fake news can be predicted with the highest accuracy using the given classifier and the value of C at 3 gives the highest accuracy. The accuracy can be altered using C which is also known as the regularization parameter. NLP played an important role in the pre-processing of the data and therefore helped the classifier to understand the data set better.

**CHAPTER # 4**

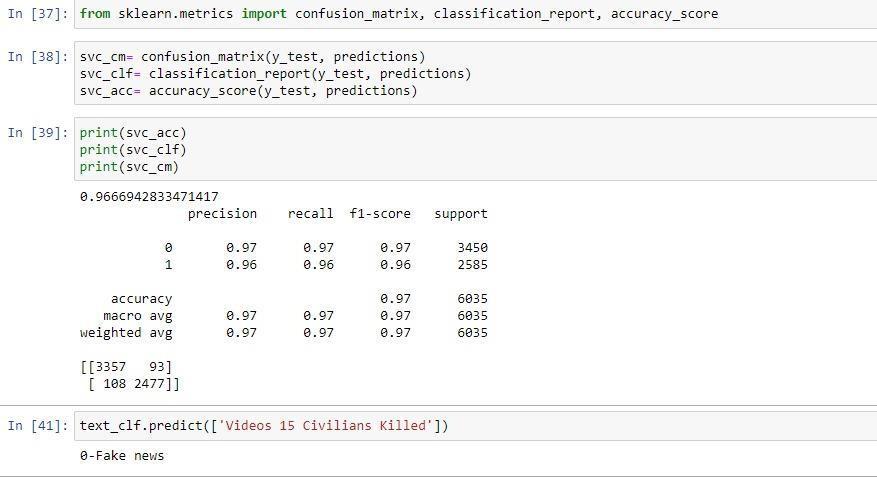
**Architecture And Design**

**1. System Architecture**

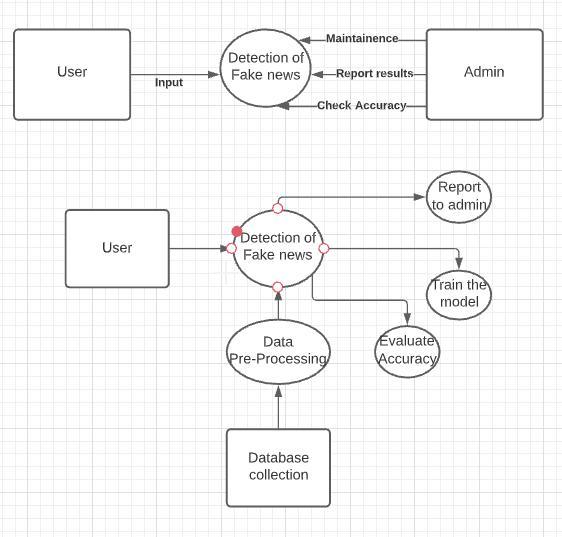


**2. Interface Prototyping**

INPUT: This is the landing page for the users. From here, users can input the news article that they wish to validate as fake or real.

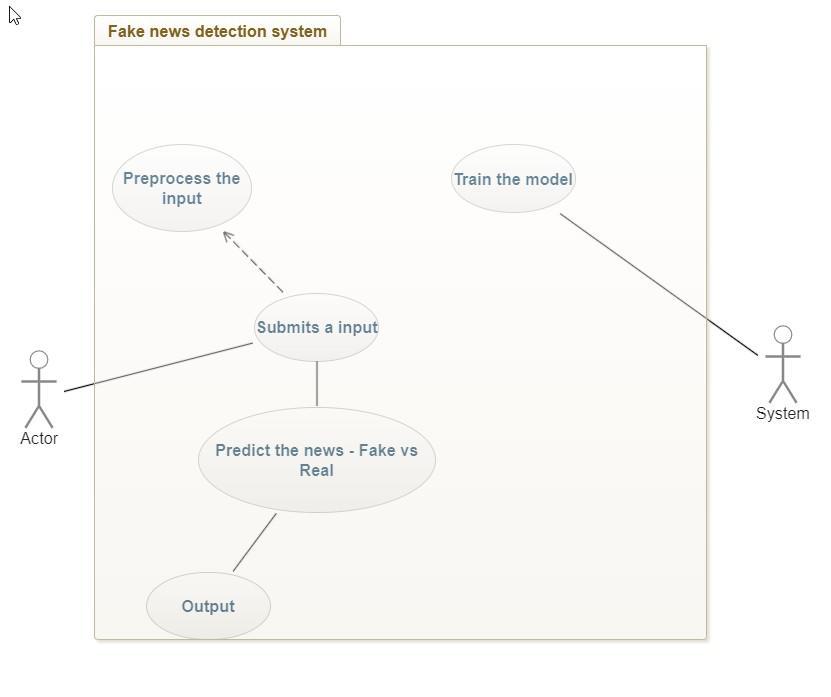


**3. Dataflow Diagram**

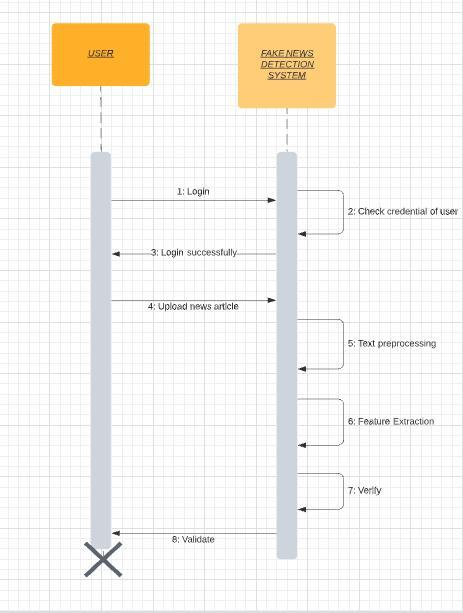


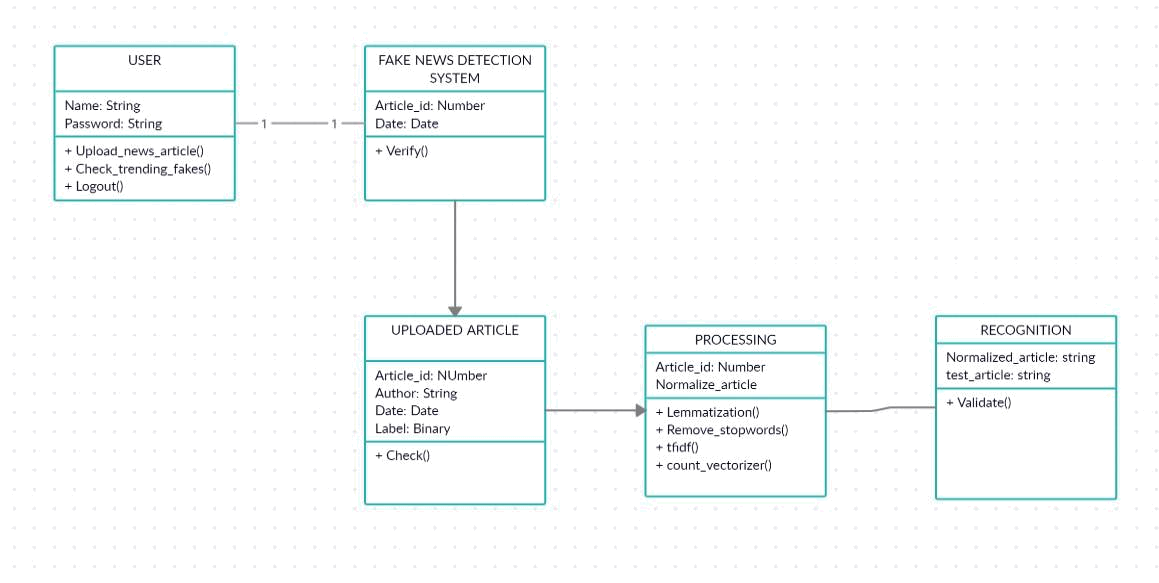
*Figure: Data-flow diagram*

**4. Use-Case Diagram**

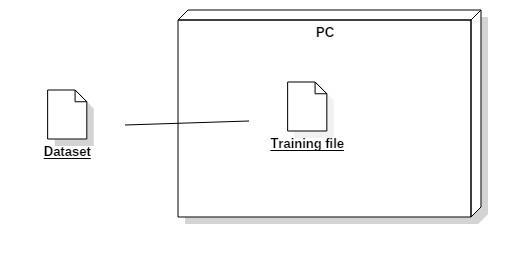


**5. Sequence Diagram**

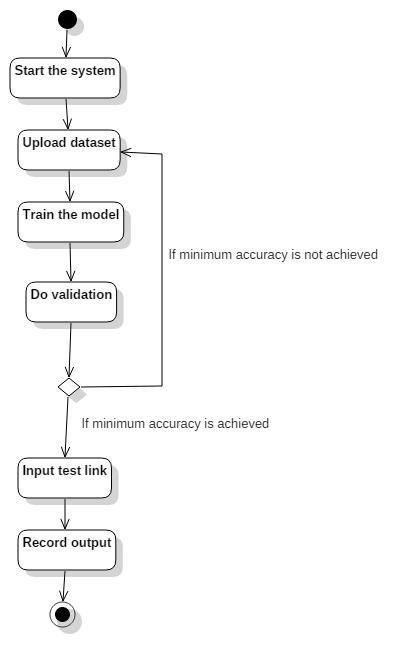




**7. Component/Deployment Diagram**



**8. State/Activity Diagram**



**CHAPTER # 5**

**Implementation**

The execution of the program is an essential stage of the project where the abstract architecture is compatible with the functional framework. The key stages of deployment are as follows:

* Planning
* Training
* System Testing and
* Transition Preparation

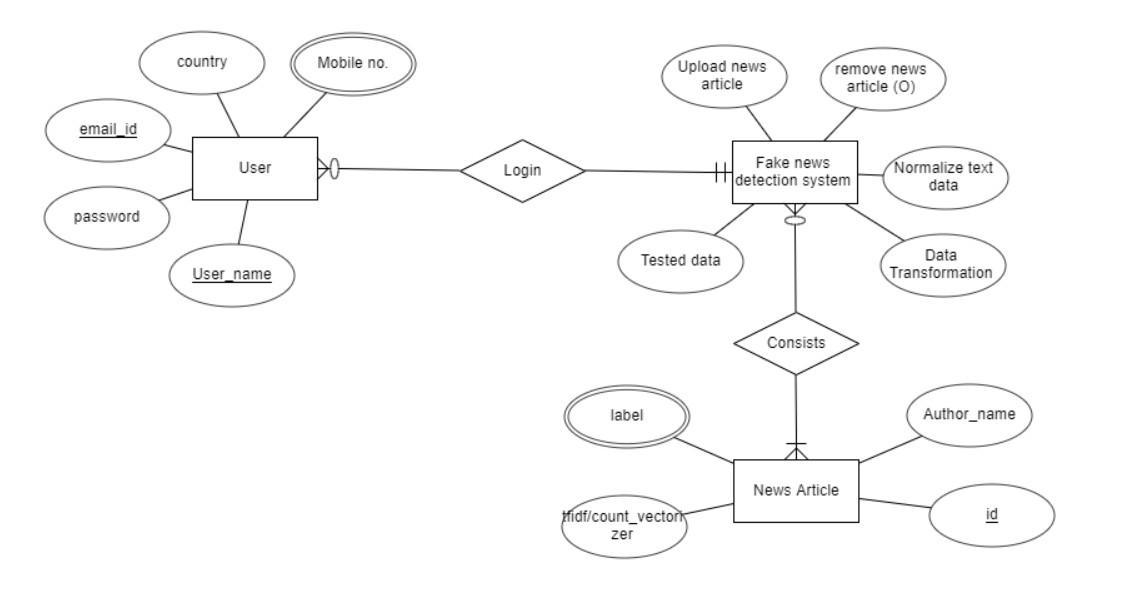
Planning is the first step in implementing the program. Planning is a recommendation on the process and period to be followed. At the time of deployment, all program workers from various agencies and system development shall be involved. They are verified by the realistic question of monitoring the various behaviors of individuals in their data processing divisions. Line managers are regulated by the implementation coordination committee. The Committee shall consider the suggestions, issues, and concerns of the customer community and shall also consider:

* Presence of the program environment;
* Self-selection and assigning of tasks for implementation;
* Consultation with unions and available resources;
* Standby facilities and communication channels;

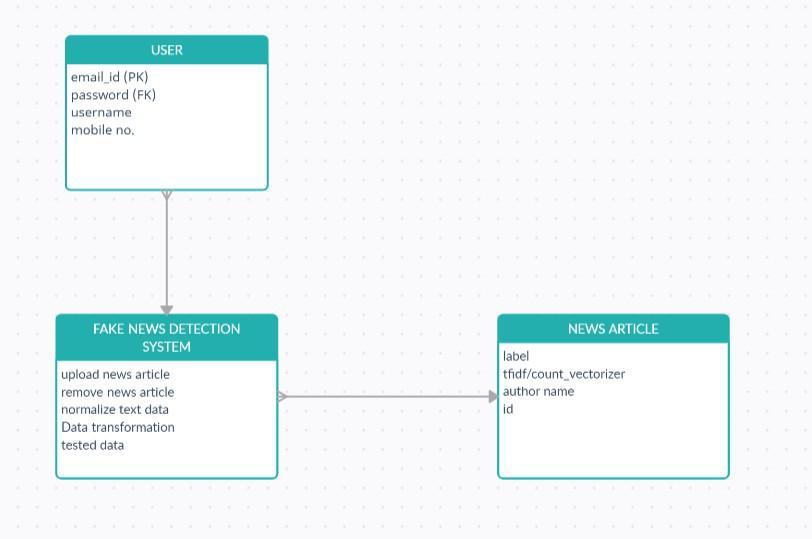
Training:

The encoding itself reflects the model's parameters of 5000 à200=1mil. Training too many network parameters requires more complex algorithms for optimization. We use the following approaches to optimize our model and to regularise it.

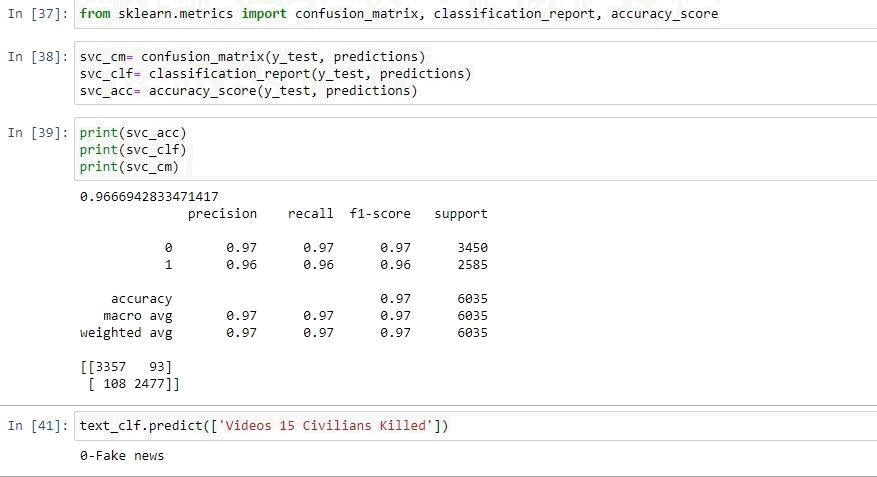
**5.1 Database Design**



**5.2 Relational Diagram**



**5.3 User Interface**



33

**5.4 Middleware**

**Anaconda**: Its goal is to simplify package management and execution. It is free and open source.

**Jupyter Notebook**: Jupyter Notebook is an open-source software framework with live programming, equations, visualizations, and text documents anytime you choose to create and upload them.

**Python with Numpy**: NumPy is a primary science programming system for Python. Among other stuff

**Python with Pandas**: Pandas is an open-source data analysis and manipulation tool that is fast, effective, scalable, and simple to use.

-Constructed upon the programming language of Python.

-A fast, efficient DataFrame object with integrated indexing for data handling**.**

**CHAPTER # 6**

**Verification And Validation**

**6.1 Unit Testing**

| TEST ID | TEST ACTIONS | INPUT | EXPECTED OUTPUT | ACTUAL OUTPUT | PASS/FAIL |
| --- | --- | --- | --- | --- | --- |
| T1 | Check the integrity of the datasets | Access the dataset abd check the attributes of the slices | Access to the dataset and all its attributes | Success | Pass |
| T2 | Checking for null values and removing them | Testing for null values | Null values replaces by ‘’ | Success | Pass |
| T3 | Creating a word cloud | Dataset of true events | Word clouds of essential words | Success | Pass |
| T4 | Implementing count vectorizer | Train datasets | Vectorized table of term frequency | Success | Pass |
| T5 | Implementing Tf idf vectorizer | Count vectorized data | Essential word vector | Success | Pass |
| T6 | Passing vectorized data into Logistic regression classifiers | Pre-processed vector | Successful prediction | Success | Pass |
| T7 | Passing vectorized data into naive Bayes classifier | Pre-processed data vector | Successful prediction | Success | Pass |

| TEST DESCRIPTION | TEST STEPS | TEST DATA | EXPECTED RESULT | ACTUAL RESULT |
| --- | --- | --- | --- | --- |
| We integrated the unit-tested module and tested the behavior complete unit. | 1) Integration of code (from units)  2) Running the code:  • Input the data sets  • Cleaning and lemmatization of the data  • Tfidf vectorization  • Passing through different classifiers to get optimal accuracy | 2 datasets, namely, test and train data sets with 20,800 entries were incorporated into the model | Efficiencies expected were as follows:  Support Vector Classifier  With C=1.0  Accuracy =95 %  With C=2.0  Accuracy= 96%  With C=10.0  Accuracy= 97% | Efficiencies obtained were as follows:  Support Vector Classifier  With C=1.0  Accuracy = 95.32 %  With C=2.0  Accuracy= 97.4815%  With C=10.0  Accuracy= 97.4816% |

***6.2 Size*-Loc**

Lines of code: 448

**6.3 Cost Analysis**

**Hardware specification for the product: -**

System: Intel Pentium. Hard Disk: 120 GB.

Input Devices: Keyboard, Mouse

Ram: 2 G

*Approx minimum cost -Rs 25000*

**Software specification for the product:-**

Operating system: Windows 7.

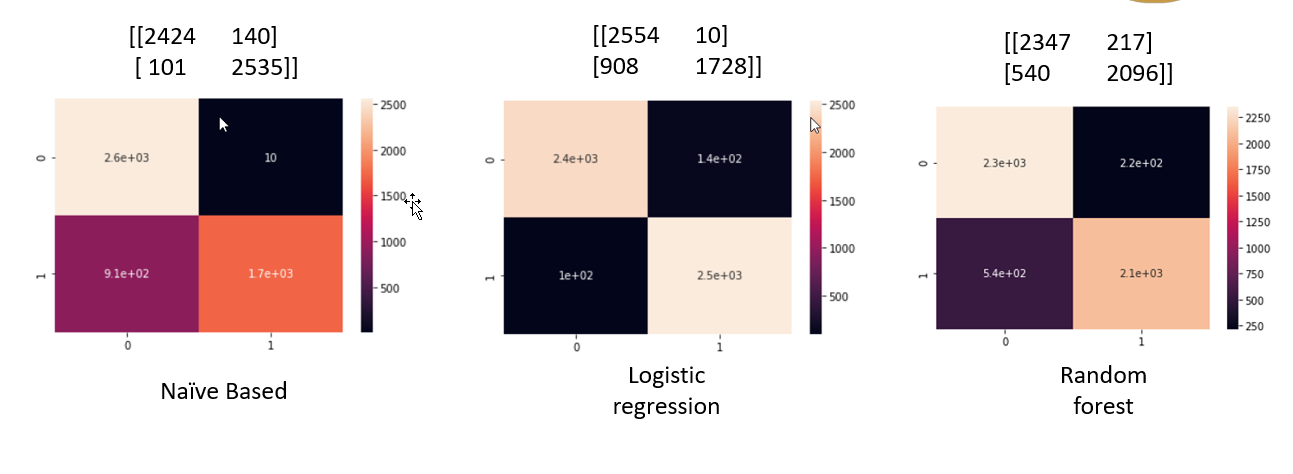
Coding Language: Python

*Approx minimum cost -Rs 5000*

***6.4 Defect Analysis***

• As a part of quality improvement and planning, every software artifact should discuss the scope of defect analysis. It helps developers to identify how issues can be prevented and in reducing or eliminate significant numbers of defects from being injected into the system. For our project purpose, we have used a confusion matrix.

• After analyzing the metrics i.e. the confusion matrix we can identify the scope of false positives and take appropriate measures to eliminate them through parameter fine-tuning or choosing a different algorithm.



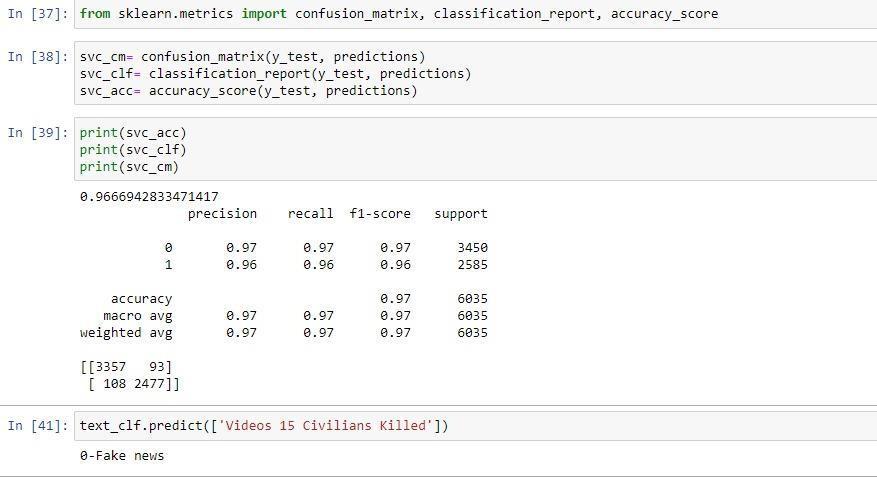
***6.5 Mc Call*’s Quality Factor**

| **QUALITY FACTOR** | **DEFINITION** | **JUSTIFICATION** |
| --- | --- | --- |
| **Correctness** | The extent to which software meets its specifications and achieves the user's mission objectives.  A software system is considered to be accurate if it meets all functional criteria | In our project, our sole objective was to classify fake news and real news in order to stop the spread the spread of misinformation.  We can accomplish this with a precision of 97% utilizing the SVM classifier |
| **Reliability** | The degree to which a programe can be trusted to fulfil its intended purpose with the needed precision is referred to as its reliability.  Reliability is a consumer impression, and faulty software might nevertheless be regarded as dependable. | We can perform our intended function with a precision score of 98% and 97% respectively for values labeled true and false using SVM classifiers. |
| **Efficiency** | The amount to which a software system uses resources such as processing power, memory, disc space, network bandwidth, and energy is referred to as efficiency.  A software should ideally use as few computational resources as feasible. | Our project is operational on  the following minimum  hardware requirements:  Intel core i5  Ram: 8GB  Hard disk space: 100GB  Which is relatively small compared to other deployed machine/deep learning modules |
| **Integrity** | The capacity of a system to withstand security threats is referred to as its integrity.  In other terms, integrity relates to the degree to which unauthorized people or programs may access software or data. | Jupyter Notebook provides provisions for setting authorization levels. A unique password can be set which restricts access to the jupyter notebook server. |
| **Portability** | The amount of effort necessary to move aprograme from one hardware and/or software environment **to another** | Jupyter Notebook and anaconda navigator are open sources. This means that anybody can share .ipynb files and run them on their respective workstations. |
| **Reusability** | To what extent portions of a software system in different applications may be utilized Reusability refers to the ability of a large piece of one product to be reused, maybe with minimal adjustments, in another product. | Our project deals with NLP so the data pre-processing steps like tokenization and lemmatization remain the same when it comes to implementation.  Those parts of the code and extracted and used for other NLP applications. |

**CHAPTER # 7**

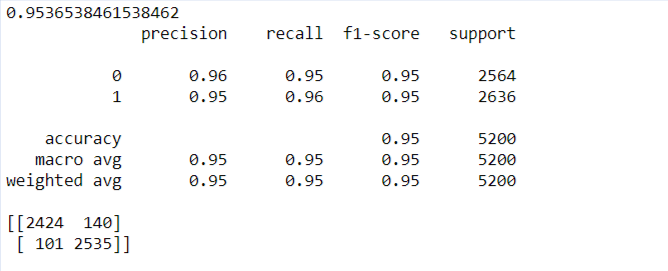
**Experiment Result And Analysis**

**7.1 Input:**

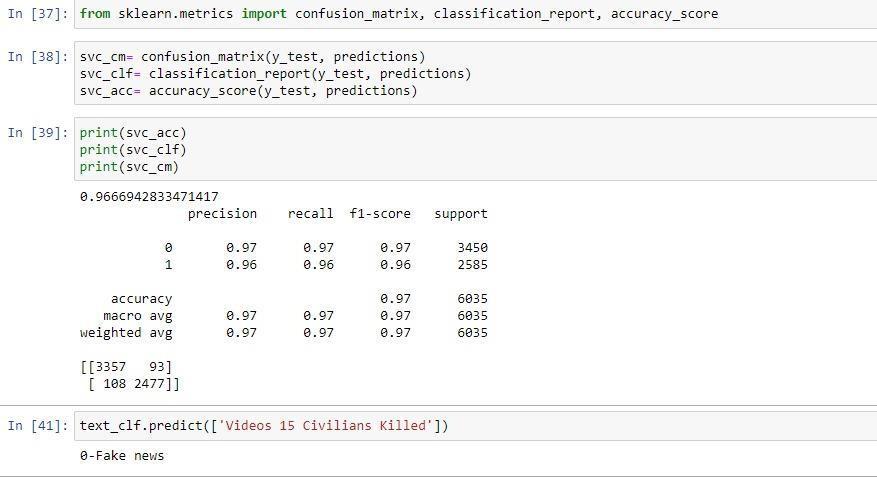


**Features of the data set:**



**Training the data:**

**Output/Prediction:**

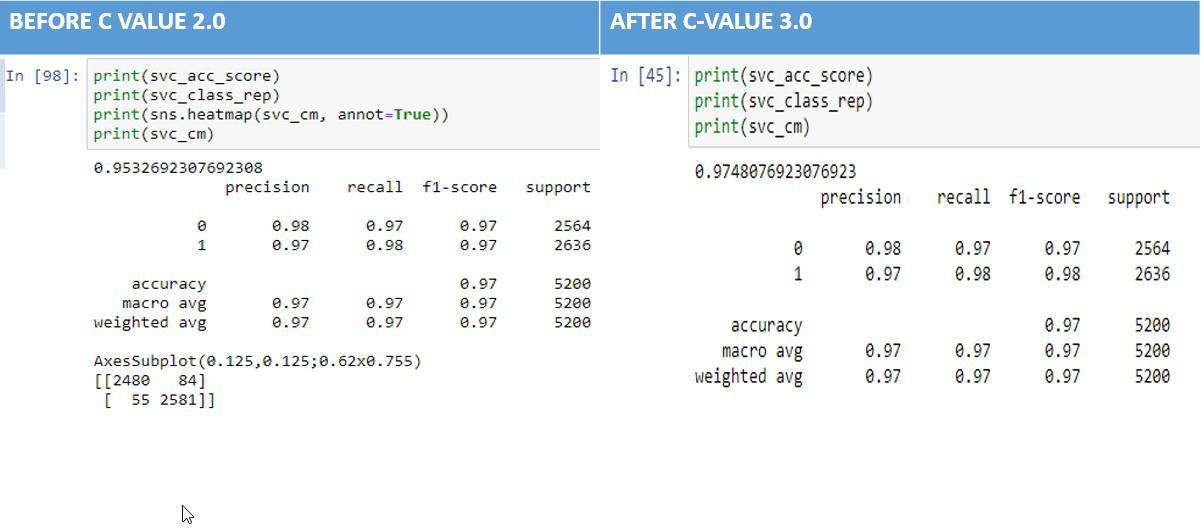


**7.2 RESULT AND ANALYSIS**

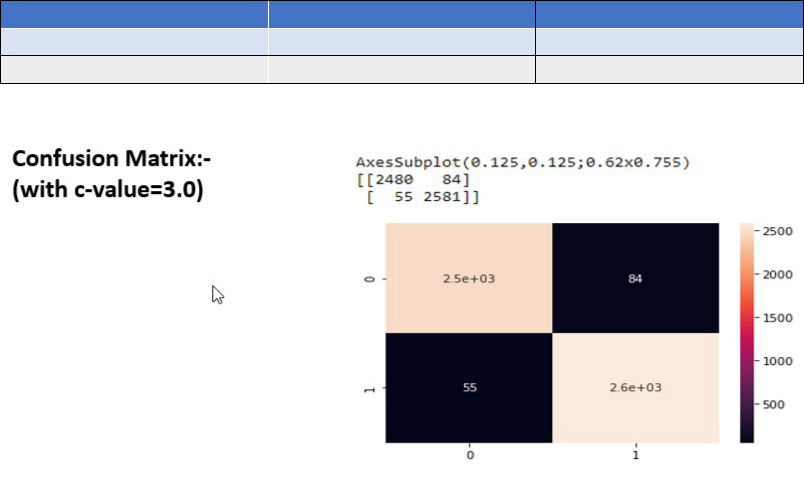
* **C value:**

Regularization parameter. The strength of the regularization is inversely proportional to C.

Must be strictly positive. The default value is 1.0.



| **Algorithm** | **C-Value** | **Accuracy** |
| --- | --- | --- |
| SVC Classifier | 2.0 | 95.32% |
| SVC Classifier | 3.0 | 97.48% |



**7.3 Conclusion And Future Work**

The model provides proof that the Support vector classifier provides higher accuracy points as compared

to Random forest, Naive Bayes, and Logistic Regression. We get this result as our model can change the C value, thereby identifying better-fit patterns with about the same computational capacity. With these findings,

we can safely infer that Fake news can be predicted with the highest accuracy using the given classifier and the value of C at 3 gives the highest accuracy. The accuracy can be altered using C which is also known as the regularization parameter. NLP played an important role in the pre-processing of the data and therefore helped the classifier to understand the data set better.

For future work findings and scope, we would like to develop ways to detect sarcasm in news articles. Sarcasm is one of the biggest challenges in NLP and every tech giant is trying their best to develop ways to detect sarcasm in text messages.

**Chapter # 8**

**References**

* **IEEE Conference 2019:** - Fake News Detection in Social Networks Using Machine

Learning and Deep Learning: Performance Evaluation.

* **IEEE Conference, 2019** 1st International Conference on Advances in Information Technology: - Fake News Detection Using Deep Learning Techniques.
* **IEEE Conference 2019**: - Fake News Detection Using Machine Learning Approaches: A systeSystematicew.
* **2018 4th International Conference on Computing Communication and Automation (ICCCA)**: - Fake News Detection Using A Deep Neural Network.