CONTEXTUAL NEWS INFORMATION RETRIEVAL

Zain Ul Abedin

Annas Israr

Faizan Ahmad

Yaldram Shahzad

Supervised by

Engr. Muhammad Umer Haroon

Submitted for the partial fulfilment of BS Software Engineering degree to the Faculty of Engineering & Computer Science

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

NOVEMBER, 2020

ABSTRACT

Data generation is an abrupt process these days, it is a challenge for different businesses to find the exact and accurate information for their statistical needs. Although this data is available for its manipulation by users. Quite amount of time is spent in retrieving this information. To gather information from various websites, data must be manually gathered. Two of the most important utilities of scraping are information retrieval for personal usage and for analytical purposes. The developed project has solved the basic challenges of personalized information retrieval. The system searches and scraps the web pages for the relevant information. Web scraping is a method that is amazingly well known and is demonstrated to have multi-domain usage nowadays. The developed project is currently limited to English news websites of Pakistan.

The tools and technologies selected for the development decide that whether the system is going to be a successful one or not. For development of this project, Scikit-Learn has been used as the main module for feature extraction and determining similarity indexes. The backend of the system is comprised of Flask and SQLite DB collectively, OCR is used to extract text from images and VS Code as the main IDE (Integrated Development Environment).

The developed system has been carefully tested using different test cases to make sure it works as expected. In this project, Gray-Box testing is used as the main technique, that is a combination of Whitebox testing and Blackbox testing. The Contextual News Information Retrieval is now working as expected and providing all the specified functionalities.

CERTIFICATE

Dated: _____

Final Approval

It is certified that project report titled '**Contextual News Information Retrieval**' submitted by **Annas Israr**, **Zain Ul Abedin**, **Faizan Ahmad** and **Yaldram Shahzad** for the partial fulfillment of the requirement of **"Bachelor's Degree in Software Engineering"** is approved.

COMMITTEE

Dr. Basit Shahzad
A/Dean Engineering Signature: _____

Dr. Basit Shahzad
HoD Software Engineering Signature: _____

Mr. Naveed Ahmad
FYP Project Coordinator Signature: _____

Engr. Muhammad Umer Haroon
Supervisor Signature: _____

DECLARATION

We hereby declare that our dissertation is entirely our work and genuine / original. We understand that in case of discovery of any PLAGIARISM at any stage, our group will be assigned an F (FAIL) grade and it may result in withdrawal of our bachelor's degree.

Group Members	**Signature**
1. Annas Israr	_____
2. Zain Ul Abedin	_____
3. Faizan Ahmad	_____
4. Yaldram Shahzad	_____

PLAGIARISM CERTIFICATE

This is to certify that the project entitled "**Contextual News Information Retrieval**", which is being submitted here with for the award of the "**Degree of Bachelors" in "Software Engineering**". This is the result of the original work by **Annas Israr**, **Zain Ul Abedin**, **Faizan Ahmad** and **Yaldram Shahzad** under my supervision and guidance. The work embodied in this project has not been done earlier for the basis of award of any degree or compatible certificate or similar title of this for any other diploma/examining body or university to the best of my knowledge and belief.

Turnitin Originality Report

Processed on 17-Sep-2020 03:28PM (UTC+0500)

ID: 1389347107

Word Count: 12389

Similarity Index

10%

Similarity by Source

Internet Sources:	01%
Publications:	01 %
Student Papers:	08%

Date: 17/09/2020 Engr. Muhammad Umer Haroon (Supervisor)

TURNITIN ORIGINALITY REPORT

Contextual News Information Retrieval by **Annas Israr**, **Zain Ul Abedin**, **Faizan Ahmad** and **Yaldram Shahzad**

From **Engr. Muhammad Umer Haroon**

Processed on 17-Sep-2020 03:28PM (UTC+0500)

ID: 1389347107

Word Count: 12389

Similarity Index

10%

Similarity by Source

Internet Sources:

01%

Publications:

01%

Student Papers:

08%

SOURCES:

1. 6% match (Student Paper)

Submitted to Higher Education Commission Pakistan

2. 1% match (Student Paper)

Submitted to Gulf College Oman

3. <1% match (Student Paper)

Submitted to University of Mauritius

4. <1% match (Student Paper)

Submitted to Texas A & M University, Kingville

5. <1% match (Student Paper)

Submitted to London School of Commerce

6. <1% match (Internet Source)

www.melchorrosario.com

7. <1% match (Student Paper)

Submitted to TechKnowledge

8. <1% match (Student Paper)

Submitted to University of Bedfordshire

9. <1% match (Student Paper)

Submitted to University of South Australia

10. <1% match (Student Paper)

Submitted to University of Northumbria at Newcastle

11. <1% match (Student Paper)

Submitted to Baze University

12. <1% match (Student Paper)

Submitted to Athlone Institute of Technology

13. <1% match (Student Paper)

Submitted to City University

14. <1% match (Publication)

Andreas Bauer, Christian Wolff. "An event processing approach to text stream analysis" , Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems - DEBS '14, 2014

15. <1% match (Student Paper)

Submitted to Abdullah Gul University

16. <1% match (Student Paper)

Submitted to Free University of Bolzano

TABLE OF CONTENTS

Chapter **Page**

LIST OF FIGURES

Figure **Caption** **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.0 Introduction

World has entered the modern era of technology and technology has captured everyone, everything, and every process in its grip. There is an enormous amount of data and information that is now available on the internet, which is available for people to access. This has made a major impact on the people's way of living and due to this almost every person is trying to get information online. The time of the users is also very precious and almost everyone wants to do their work in a more convenient way. There are about billions of articles and newspapers with numerous facts and important information. To look for some desired facts or news a person needs to go through multiple sources online to get their required information. So, to solve or minimize this problem a web-based application is developed for contextual extraction of relevant news information according to the required query that was entered. By using this application, the user is be able to search the desired news that is automating the process of searching and extract the required information for the user which would save a lot of time. This app also merges the extracted information from different sources in a document along with the references that the user would be able to import in its local device.

1.1 Motivation

As there are tons of information over the internet, sometimes searching for your required news/information can be hectic and frustrating at the same time. The motivation behind the development of this application was to reduce wastage of time and effort that people had to face to get the relevant news of their interest and secondly, to provide news from authentic sources.

1.2 Problem Statement

In the modern world of technology internet is available to almost everyone. Internet is now the biggest means of mass media and by using this one can change the ideology and way of thinking of a huge number of people. Authentic and valid news always fade away with paid promotions and thus users are unable to find the right news from right source. Different type of fake news and propagandas keep circulating all around the internet. This system extracts your required information from trusted and authorized resources with links and references of these resources automatically and thus saving a lot of time and effort of the user.

1.3 Goals and Objectives

Crucial goals that serve the application for users are to register user, search news, show newsfeed, show latest news, save news information, select user interests from predefined interests (Sports, Entertainment, Business, Pakistan), and show news search history of the users.

This project assists those that are linked with news research including most of analysts, anchors, reporters. This project retrieves valid and authentic news information from sites, articles and blogs available on internet according to users context. This project then manages the gathered information by displaying information in an appropriate way to support the user in an efficient way.

1.4 Scope of the study

Table 1.1 Scope of the study

Scope of the study	
Title	**Contextual News Information Retrieval**
Description	A web-based application that automatically searches and extracts the news from different sources on the internet based on the users query and provides the user with a merged document of the information along with the links and references.
Justification	The main aim is to save user's time of browsing over internet for desired information and to make sure that the news is coming from authorized and trusted sources.
Constraints	Limited number of Pakistani English newspapers for Latest news and Newsfeed.
Assumption	A fast & stable internet connection is available, a working web browser.
Stake holders	Users, news researchers, news reporters, developers, testers, admins.
Risks	Lack of internet, slow internet connection, faulty internet browser.
Deliverables	A web-based application, final documentation report.

1.5 Process Model

In the development of this system the incremental process model is used. The work is broken down into different modules. The modules were then developed integrated and tested. Different aspects of the system are requirement gathering, work breakdown and decomposition, designing, coding, integration, testing and deployment.

1.5.1 Requirement gathering

The requirements and specifications that were needed for Contextual News Information Retrieval system were gathered from different sources.

1.5.2 Work breakdown and decomposition

The system was decomposed into the modules in this phase. These modules include scrapping of texts, image processing for conversion into text, scrapping data on images, extracting data from sources, extracting data according to the interest of the user, merging data at one place, making data downloadable for users on their local machine and providing references of the extracted news information.

1.5.3 Designing

In this phase, the different requirements of the system are transformed into properly defined high level specification of each of the modules of the system.

1.5.4 Coding

The actual development and coding according to the requirements and design specification has been done in this phase and has been repeated for each module of the system.

1.5.5 Integration

After the different modules were developed, they were now integrated at one place in this phase.

1.5.6 Testing

Different modules have been tested parallel to the development, but the main system testing, and high-level testing has been done in this phase after the integration of the modules.

1.5.7 Why incremental model

Incremental model has been used because, tasks can be divided, problems can be easily identified and managed, less skilful team can manage it, can focus on the module under development.

1.6 Nature of the project

Contextual news information retrieval is a web-based application that is made using python at the back end for data mining and machine learning algorithms like Image processing and Web Scrapping. Flask SQLITE DB have been used for maintaining the database.

1.7 Overview of the Report

In this chapter, the surface details of Contextual News Information Retrieval have been discussed. What is the system that has been developed in highlighted, why it has been developed, how it's being operated, how it is impacting on the modern world, for whom it has been made, for which domain and field it benefits the most, where has it been organizationally located and where will it be in the future.

CHAPTER 2

BACKGROUND AND EXISITING WORK

2.0 Introduction

This chapter mainly focuses on existing systems that are related to CNIR (Contextual News Information Retrieval). News researchers and students of journalism spend a handsome amount of their time in researching about various topics, under their consideration. CNIR provides a solution to save their time and effort. They can just simply search and leave rest of the work to the system, CNIR searches their related information and retrieves the information related to their query. Google News is an existing system which does a pretty similar job, but it provides you links, and you have to individually go through those links to get your desired information and it contains only articles and blogs. ABC News is another system which provides the link to sources and does not have any mechanism to download the information on your system. CNIR is solving these problems.

2.1 Explanation of Important Constructs of the Application Domain

The techniques that are used and play vital roles in the given system are briefly explained as following.

2.1.1 Query

When a user comes to an information retrieval system, he has some question in mind. Those questions are then transformed into a query which he writes and expects the system to answer that.

2.1.2 Information Provision

As the name suggests, the application is assumed to fetch the information that user has asked in its query. Web Scraping refers to getting data from webpages by writing scripts. These scripts automatically get the data from the page's html when executed. Web scraping is used for providing the users with their desired information.

2.1.3 Information Representation

Information provision is another important construct. Information when not presented in a suitable format, has no impact on its reader.

2.1.4 Document Similarity

When providing information according to the context of user, it is particularly important to make sure the provided information relates to the entered query. This is made sure by finding the document similarity, which means by finding out the similarity between the entered query and information. The content which has the maximum similarity with reference to the query is then provided to the users.

2.2 Existing Systems

Existing systems relevant to CNIR are listed below.

2.2.1 Google News

Google News is a news aggregator application to search different news made by google. It provides the facility of daily updates, provides search facility, it provides different categories for the users to search information. But its drawback is that it only provides you with links and you have to individually visit those links and there is no mechanism to download the information [1].

2.2.2 ABC News

ABC News is another news website by American Broadcast Company. It provides searching facility and allows you to see latest stories. But again, it does not have any mechanism to download information and it does not analyse pictorial information and E-Papers [2].

2.3 Comparison of Existing Systems

A comparison between two existing systems is given below in Table 2.1.

Table 2.1 Comparison of existing systems

Features	**Google News**	**ABC News**
Text Extraction	No	No
Text Extraction from Pictures	No	No
Analyzation of user's script	No	No
Download Option	No	Yes
Searching Option	Yes	Yes

2.4 Summary

In this section, summarization of some existing systems for news search is discussed which are Google News, ABC News etc but they all lack at some point i.e. data representation or download option etc. The system solves these problems and tackles these features. In this way, the system facilitates its users and provides them ease and convenience.

Most importantly, the system saves the time of its users and searches the information on their behalf rather than letting them go through the individual links themselves, and then the users can download and keep the results with themselves which further provides convenience.

CHAPTER 3

REQUIREMENTS SPECIFICATION

3.0 Introduction

This chapter mainly focuses on system requirements of the system. System Requirements are those software or hardware requirements that cause the system to do what it is supposed to. The absence of these requirements can cause serious system failure, system failure can be in the form of performance issues or compatibility issues. Performance issues are those which cause system to perform below the bar i.e. the system can hang, or crash and compatibility issues can be when the deployed system is not compatible with the working environment or the hardware installed at the workplace. System requirements of the application are discussed in this chapter. System requirements of the system include interface requirements, database requirements, resource requirements, software interface requirements, hardware interface requirements, functional and non-functional requirements. Apart from these system requirements the use cases of the system are also discussed in this chapter.

3.1 Interface Requirements

Those necessary measures which need to be met in order to interface different components of the system are termed as interface requirements, as the system has many different modules and components so interfacing all these separate components is a crucial part in getting the required functionality. In this section, the need to discuss that what is needed to make the system communicating with all its components. As the system is a web-based application that is why the interface requirements are subdivided into software interface requirements and hardware interface requirements.

3.1.1 Software Interface Requirements

As the system can be accessed online, the software interface requirements conclude that the user must be having an active network connection and google chrome/Mozilla Firefox (web browsers) etc.

3.1.2 Hardware Interface Requirements

These requirements include minimum hardware resources required to run a suitable operating system. In every scenario, user wants to assurance that their system resources are enough to support their tasks and perform all the required functions and services like running applications on the server with ease. As the system is an online web app, it can be easily accessed on a system with a 1.8 GHz dual core processor, 4GB of ram and 500 GB of disk space.

3.2 Functional Requirements

The requirements which are clearly specified in the phase of requirements gathering are functional requirements. These are must for any system to be successful, functional requirements are the critical requirements that deal directly with the core functionality of the system so these must be met by the system.

3.2.1 User Registration

System should have the capability to register a new user (sign up) by prompting its name, email, and a password. System should have the capability to authenticate registered user (sign in) by prompting the email and a password. System should be able to give control to users for updating their account name and account password.

3.2.2 News Search

System should be capable to analyse the users input to determine the users context. System should enable the user to search news of its choice. System should be able extract news information available online. System should be able to extract news information from newspaper images.

3.2.3 Display of Information

System should be capable of displaying the extracted information in a readable form. System should display the information based on user context.

3.2.4 Latest news

System should be able extract latest news information from The Nation, Dawn, Pakistan Today, Daily Pakistan and Express Tribune.

3.2.5 Newsfeed

System should be able to show newsfeed based on the search history and interest of registered users from The Nation, Dawn, Pakistan Today, Daily Pakistan and Express Tribune.

3.2.6 Search History

System must save search history of the registered users. System must update news results against search history of registered users to make searching capabilities more efficient.

3.2.7 News Interests

System should enable the users to select news interests from sports, entertainment, business, and Pakistan categories.

3.2.8 Save information on local machine

System should be able to give control to user by providing option to save and export the news information in user's device.

3.3 Non-Functional Requirements

Requirements which are not clearly stated during phase of requirements elicitation are known as non-functional requirements. These are some general/common needs and are considered to improve the quality of the system. These requirements are not directly related to the system's functionality but add more clarity and a greater user experience to the users. The more often they are implemented the more quality the system can have. Almost all systems have these kinds of requirements and it has been made sure that these requirements are implemented where necessary.

3.3.1 Security

Server end data shall be accessible to developers/administrators. The IDEs and modules used in development of this application also provides built-in security by itself. Sensitive data will be encrypted before being sent over insecure connections like the internet. Other security measures are also kept in mind.

3.3.2 Reliability

System shall provide the database and backup of database to ensure the reliability of overall system. Since the DB is the main aspect of the application so it shall be maintained, updated, and secured along with all reliable components of system. Regular backups have been taken to insure reliability.

3.3.3 Availability

Availability of the system shall be determined to be available anytime, anywhere to its audience whenever the users need it. However, the time when the server or database is down or there is some technical difficulty, the performance and availability of the system shall be compromised. Any possible backup shall be available and shall be provided to users in this case until the problem has been identified and corrected. It is made sure that no such problem occurs, and all exceptions are handled responsibly.

3.3.4 Maintainability

System shall be maintainable and is maintained regularly to ensure that new changes can be implemented. Flask SQLITE Database is used to maintain the database and user data, server is used to maintain the database and to look after the overall application. The

downloaded information is saved inside a file, which is available for users to download on their system. So, it is easy to expand the user data.

3.3.5 Portability

The system is an online web app, so end user part is fully portable, and it shall run on every web browser and platform including the most popular ones to ensure the portability and compatibility related to every browser available. The popular platforms include Windows, Linux, and MAC OS. Common browsers including Mozilla Firefox, Google Chrome, Microsoft Edge etc all are compatible with the application and can show all the available content without any issues.

3.4 Use Case Model

This is a model that is designed based on user goals. It serves as communication interface that shows the interactivity between user and application, response of application towards user [3]. Use Case diagrams are a comprehensive way to show the interaction between different user roles and the system. This system also has a main use case diagram and written use cases with diagrams that has many functions for the user to do. The system shows a homepage or landing page to the guest user that has many sub features related to website. The home page has a search bar from where users can search news of its own choice and context. Then in that same home page there is also a feature of latest news from where user can see the trending news. Apart from these guest user features there are some essential features that are meant to be used and performed by registered users only. These features include user registration, newsfeed, search news, saving of news, news interest selection, viewing search history. These interesting features makeup the system named as Contextual News Information Retrieval (CNIR). The below use case diagram shows how a guest user and a registered user can access system functionalities and the difference between their roles as a user. A detailed view of Use Case diagram for Contextual News Information Retrieval is given below in Figure 3.1:

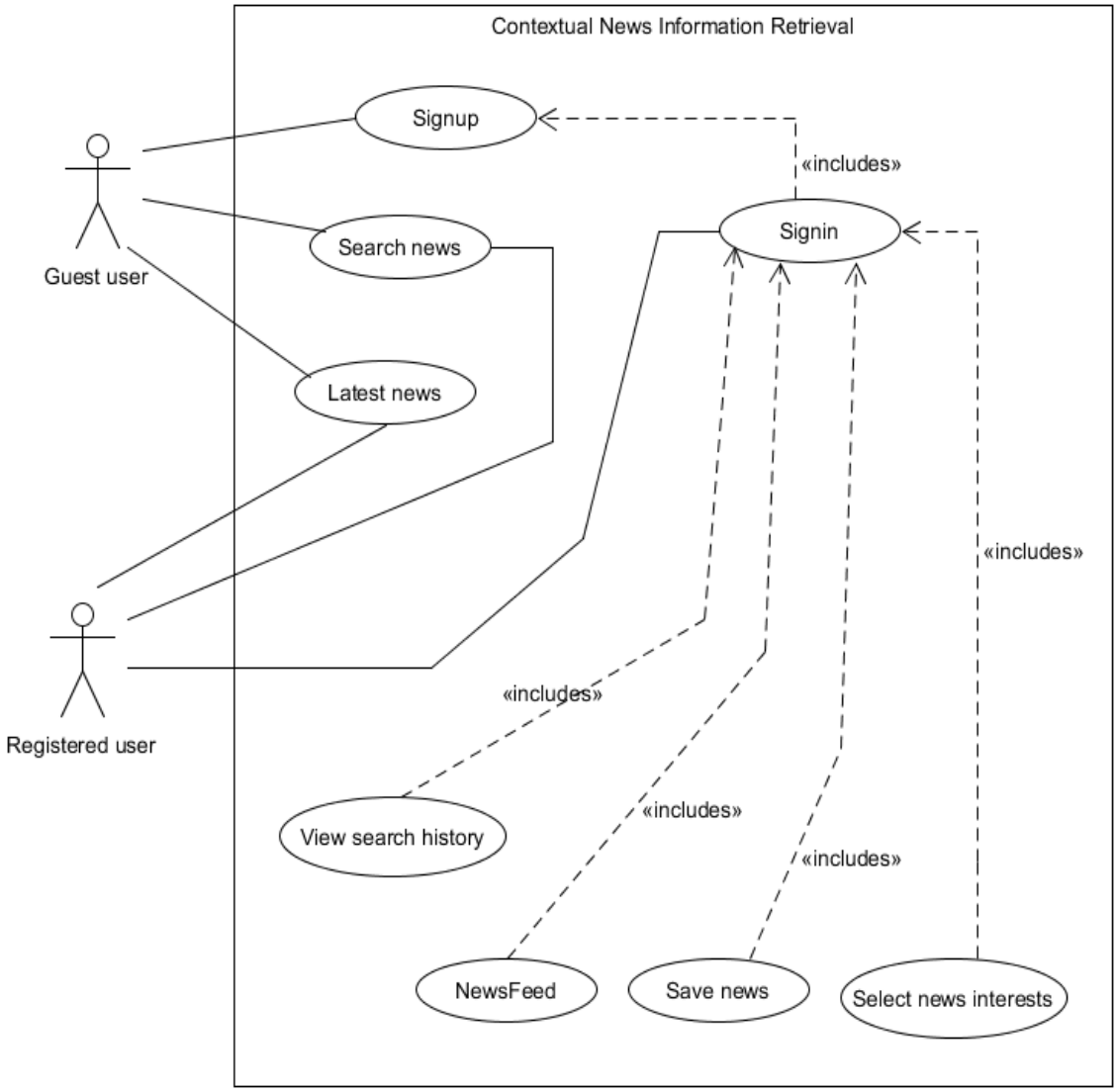

Figure 3.1 Use case diagram for Contextual News Information Retrieval

3.5 Use Cases

Primary use cases that serve purpose of user goals sign in, sign up, latest news, newsfeed, select news interest, save news, search news, and view search history. There use cases are stated below:

3.5.1 Sign in

In figure 3.2, the user is asked to input the login credentials and click sign in. After the user input, the system in response to that input verifies the email and password and passes the user to more feature rich home page to unlock and use special features. The system also gives response if the sign in gets unsuccessful due to some reason and shows an appropriate message.

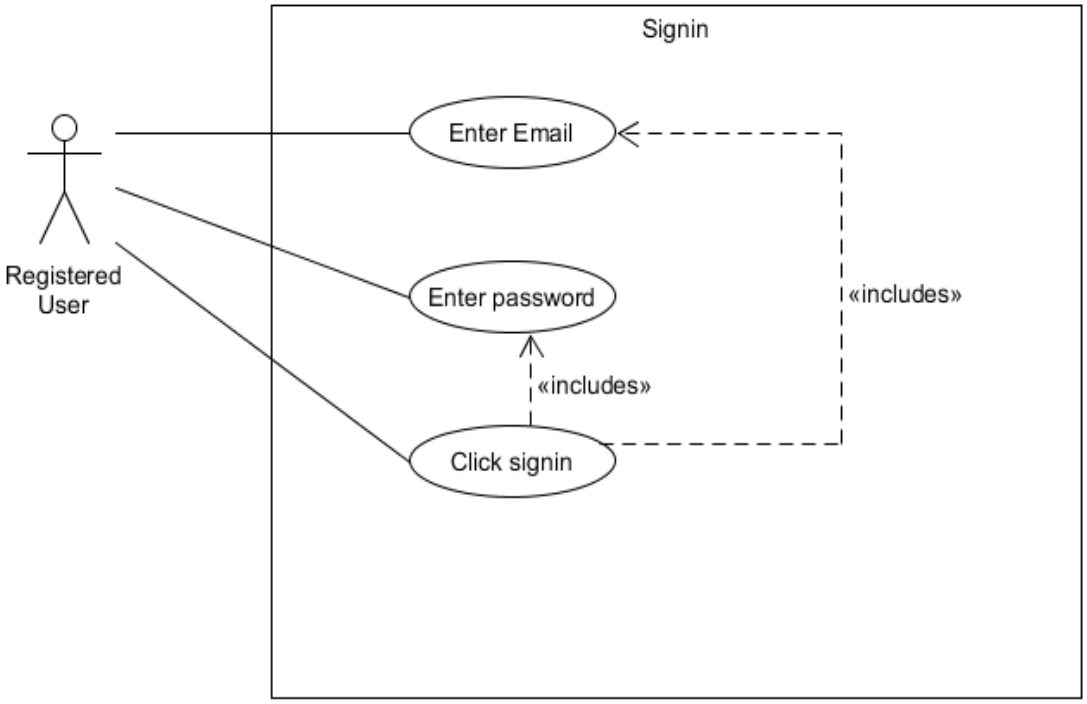

Figure 3.2 Use case diagram for sign in

Use case for Figure 3.2 is given below in Table 3.1.

Table 3.1 Use case for sign in

Use case ID	UC-1
Use case name	Sign in
References requirement	Requirement no. 1
Actors	Registered user
Purpose	To sign in registered user
Overview	The user enters the email and password and then clicks on sign in
Type	Primary and essential
Pre-Condition	User must be signed up
Post-Condition	User signed in successfully
Normal flow	

Actors Actions	System Response
1. The user enters the email and password that was entered during signup.	2. The system verifies the user email and password and shows next phase.
Alternative flow	
Step 1: The user enters the wrong password.	Step 2: The system does not pass the user to next phase and shows a message of incorrect password.

3.5.2 Sign up

In figure 3.3, the system prompts the guest user to enter first name last name, email and password and click sign up. In response to that the system validates the details and shows an appropriate message on successful sign up. The system also gives response if the sign up gets unsuccessful and shows an appropriate massage.

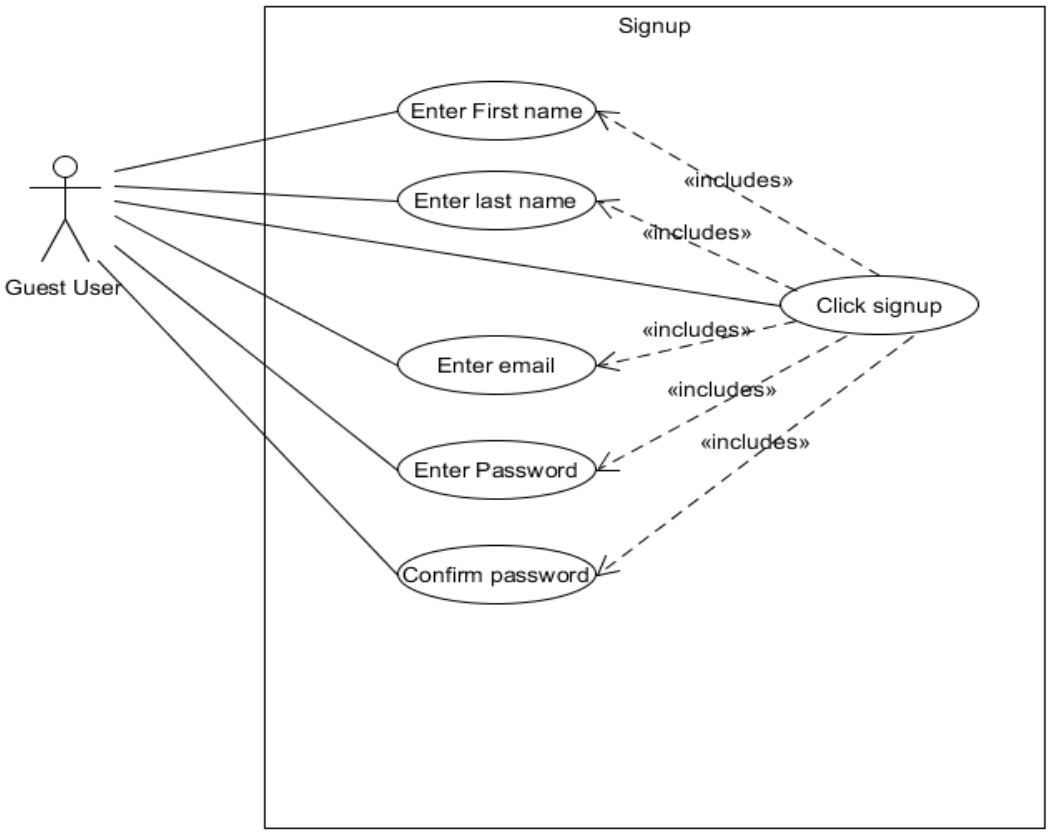

Figure 3.3 Use case diagram for sign up

Use case for Figure 3.3 is given below in Table 3.2.

Table 3.2 Use case for sign up

Use case ID	UC-2
Use case name	Sign up
References requirement	Requirement no.2
Actors	Guest user
Purpose	To sign up the guest user
Overview	The user enters the first name last name, email, password, password and then click on sign in
Type	Primary and essential
Pre-Condition	None
Post-Condition	User successfully signed up
Normal flow	
Actors Actions	System Response
1. The user enters the first name, last name, email, password and confirms the password.	2. The system creates an account against the user inputs and pass it to next phase.
Alternative flow	
Step 1: The user enters the wrong password in confirm password field.	Step 2: The system does not pass the user to next phase and shows a message of mismatch password.

3.5.3 Search news

In figure 3.4, the system prompts the user to enter the keywords and click search. In response to that the system analyses the user's context and shows the news results in an

appropriate way. The system also gives response if keywords are too short to be use as context and shows an appropriate massage.

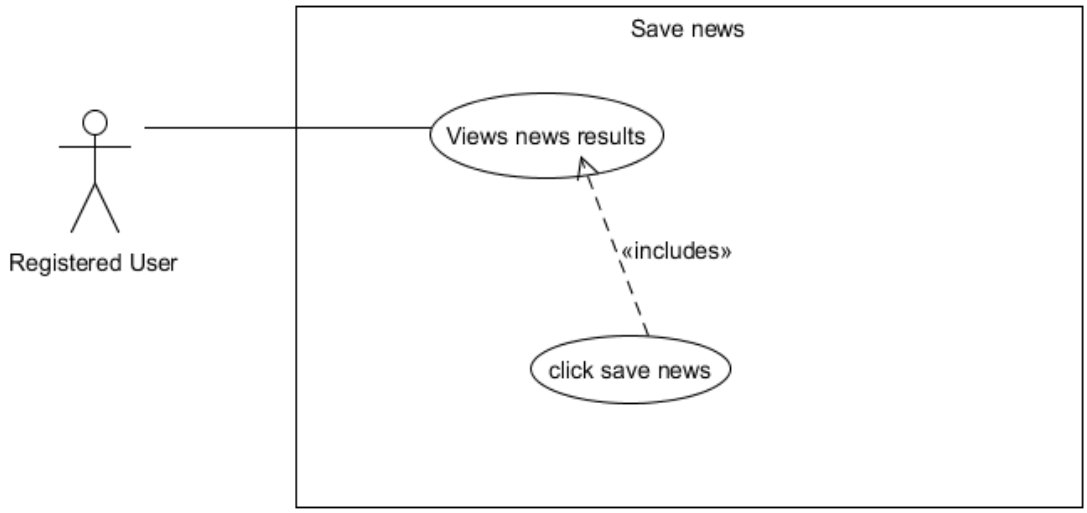

Figure 3.4 Use case diagram for search news

Use case for Figure 3.4 is given below in Table 3.3.

Table 3.3 Use case for search news

Use case ID	UC-3
Use case name	Search news
References requirement	Requirement no. 3
Actors	Guest User, registered user
Purpose	To search news
Overview	The user enters the keywords and then clicks on search
Type	Primary and essential
Pre-Condition	None
Post-Condition	Search successfully performed
Normal flow	
Actors Actions	System Response

1. The user enters the keywords.	2. The system passes the keywords to find results.
Alternative flow	
Step 1: The user does not enter anything.	Step 2: The system shows a message to enter some keywords.

3.5.4 View search history

In figure 3.5, the application displays search history of the registered user and allows the registered user to manage it by giving an option to clear the search history. The system gives response on successful deletion of search history and shows an appropriate massage. The system also gives response on unsuccessful deletion of search history and shows an appropriate massage.

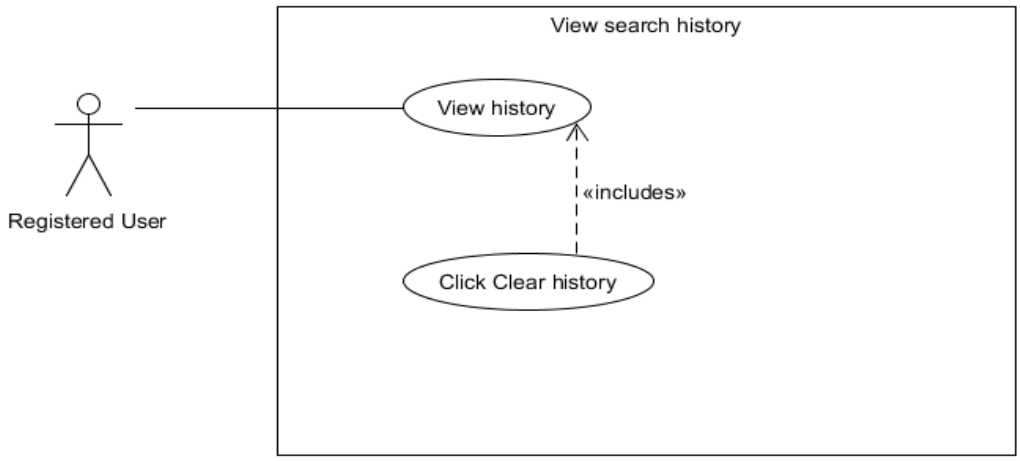

Figure 3.5 Use case for diagram view search history

Use case for Figure 3.5 is given below in Table 3.4.

Table 3.4 Use case for view search history

Use case ID	UC-4
Use case name	View search history
References requirement	Requirement no. 4
Actors	Registered User

Purpose	To show search history to registered user
Overview	The user sees search history and later clicks on delete history to delete history
Type	Primary and essential
Pre-Condition	User must be signed in
Post-Condition	Search history successfully managed
Normal flow	
Actors Actions	System Response
1. The user clicks on delete history.	2. The system shows message that search history deleted.
Alternative flow	
Step 1: There is no history to delete and user clicks delete history.	Step 2: The system shows message no history found to be deleted.

3.5.5 Select news interests

In figure 3.6, the system shows predefined news interest to registered user and allows the user to update the news interest by choosing any one of them and save them. The system gives response updating of news interest and shows an appropriate massage. The system also gives response on updating of news interests and shows an appropriate massage.

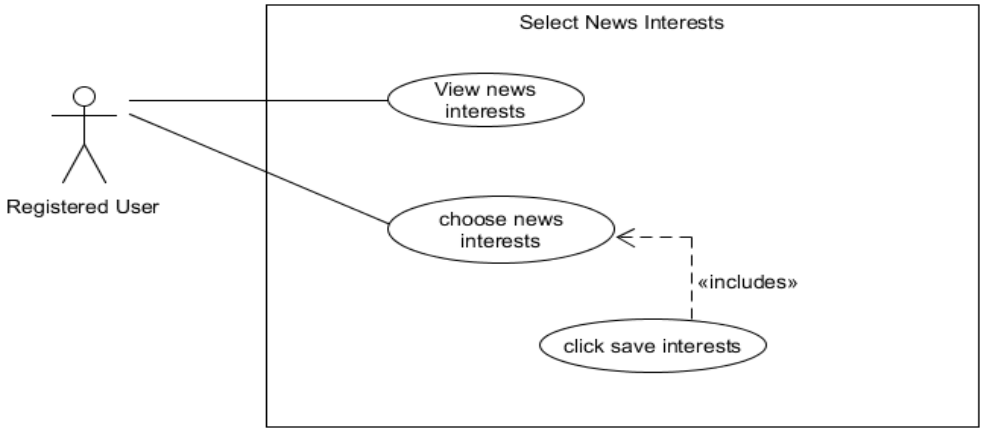

Figure 3.6 Use case diagram for select news interests

Use case for Figure 3.6 is given below in Table 3.5.

Table 3.5 Use case for select news interests

Use case ID	UC-5
Use case name	Select news interests
References requirement	Requirement no. 5
Actors	Registered user
Purpose	To choose and save news interest of registered user
Overview	The user sees the news interests and choose the predefined interests and clicks save
Type	Primary and essential
Pre-Condition	User must be signed in
Post-Condition	News interest successfully selected
Normal flow	
Actors Actions	System Response
1. The user sees the news interests to chooses interests and then the user clicks save to save news interests.	2. The system shows news interests saves news interests selected by user.
Alternative flow	
Step 1: The user does not select any news interests and clicks update interests.	Step 2: The system does not save news interests.

3.5.6 Save news

In figure 3.7, the system gives option to save news in users local machine to registered users. The system shows response by creating a file containing news result and make it downloadable.

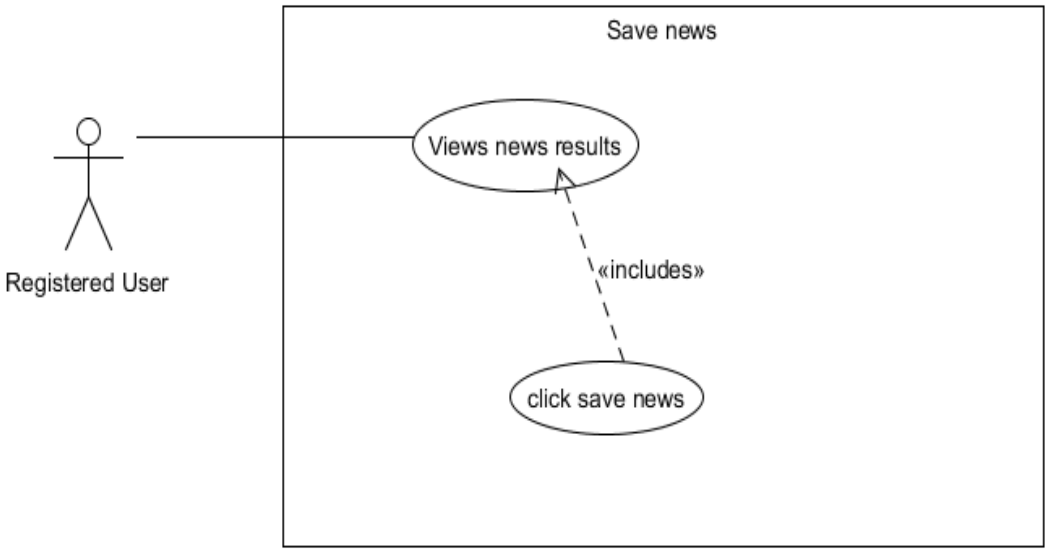

Figure 3.7 Use case diagram for save news

Use case for Figure 3.7 is given below in Table 3.6.

Table 3.6 Use case for save news

Use case ID	UC-6
Use case name	Save news
References requirement	Requirement no. 6
Actors	Registered user
Purpose	To save news
Overview	The user sees the news information results and later clicks on save news to download locally
Type	Primary and essential
Pre-Condition	User must be signed in

Post-Condition	News successfully saved
Normal flow	
Actors Actions	System Response
1. The user sees the news results and clicks save news.	2. The system saves the news results in users device.
Alternative flow	
Step 1: The user sees the news results and does not clicks save news results and leaves the page.	Step 2: The system shows a message to ask if the user wants to save the search news results.

3.5.7 Latest news

The figure 3.8 shows latest news to the users. The system responds by showing the latest news and provides appropriate feedback on successful result of latest news.

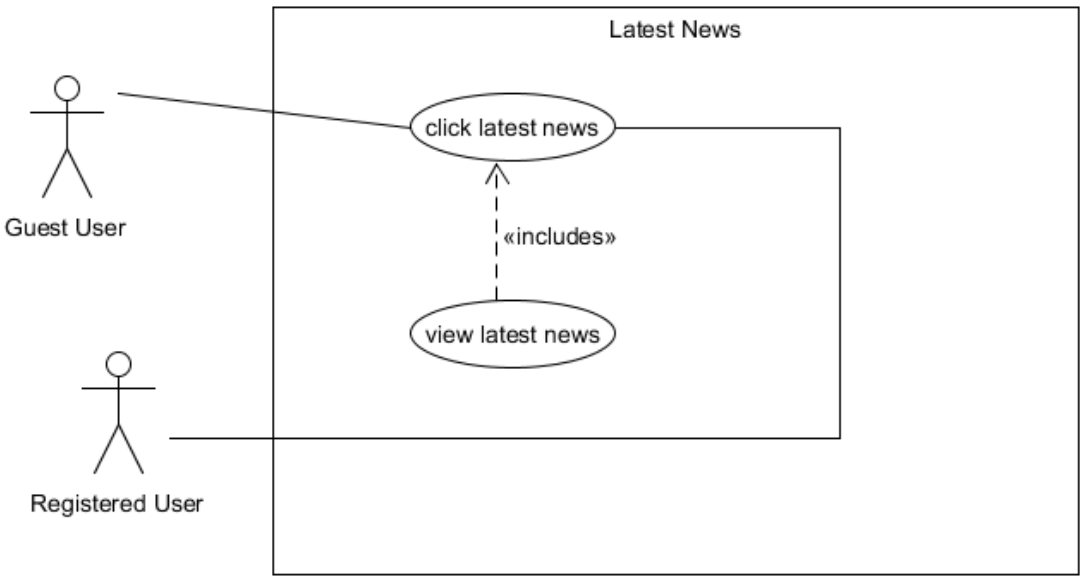

Figure 3.8 Use case diagram for latest news

Use case for Figure 3.8 is given below in Table 3.7.

Table 3.7 Use case for latest news

Use case ID	UC-7
Use case name	Latest news
References requirement	Requirement no. 7
Actors	Registered user, guest user
Purpose	To show latest news
Overview	The user clicks on latest news and then system shows latest news to the respective user.
Type	Primary and essential
Pre-Condition	Users must be active
Post-Condition	Latest news successfully shown
Normal flow	
Actors Actions	System Response
1. The user clicks on latest news and see the latest news	2. The system shows the latest news to the users.
Alternative flow	
Step 1: The user clicks on latest news button but does not see latest news	Step 2: The system shows the message about unavailability of latest news.

3.5.8 Newsfeed

This news case shows newsfeed to the users. The system responds by showing the newsfeed and provides appropriate feedback on successful result of newsfeed.

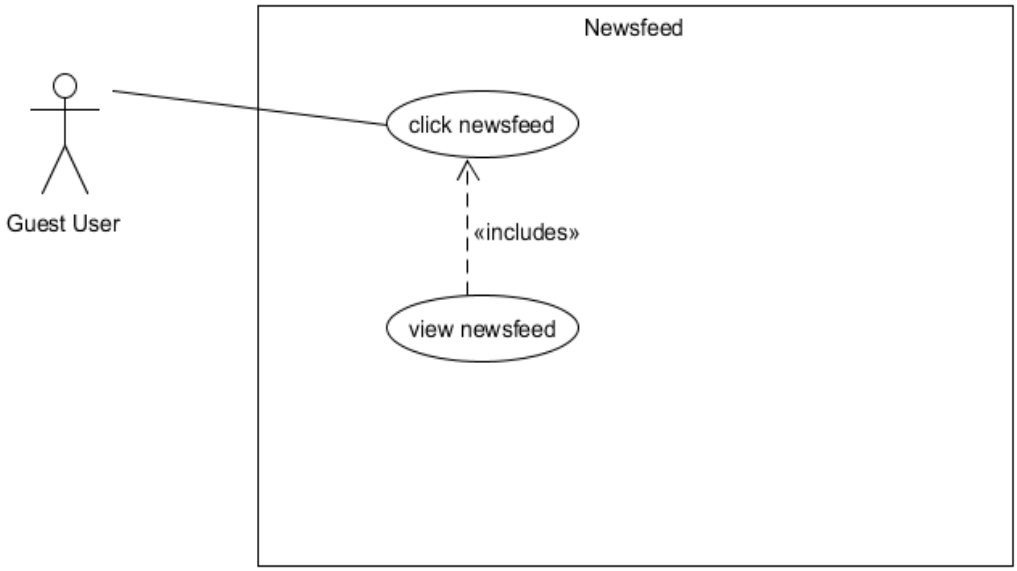

Figure 3.9 Use case diagram for newsfeed

Use case for Figure 3.9 is given below in Table 3.8.

Table 3.8 Use case for newsfeed

Use case ID	UC-8
Use case name	Newsfeed
References requirement	Requirement no. 8
Actors	Registered user
Purpose	To show newsfeed
Overview	The user clicks on newsfeed and then system shows newsfeed to the registered users.
Type	Primary and essential
Pre-Condition	User must be signed in
Post-Condition	Newsfeed successfully shown

Normal flow	
Actors Actions	System Response
1. The user clicks on newsfeed and see the latest news	2. The system shows the newsfeed to the users.
Alternative flow	
Step 1: The user clicks on newsfeed button but does not see newsfeed news	Step 2: The system shows the message about unavailability of newsfeed.

3.6 Resource Requirements

These are the items that are used as resource in software and hardware to achieve a task. These tools are used in the project:

3.6.1 Web Scraper

Web scraper is data extraction tool to extract data from websites. It helps in extracting data from web pages, tables, html pages etc. It uses algorithms of different python frameworks, packages, and libraries to scrap data. Logics can be made to scrap data of specific choice and context.

3.6.2 Tools

Here are some languages, frameworks, architectures that are used as tools to develop the system.

HTML: It is a mark-up language; it is used to create structure of webpages.

CSS: Cascading Style Sheet is a language that is written within HTML to design and decorate webpages to look better.

Bootstrap: Bootstrap makes the handling of CSS easy and makes the pages responsive on different screen.

JavaScript: It is a scripting language that is used to handle and control the behaviour of web pages on web.

Python: It is a high-level language that is used at frontend and backend as well. It is used because of its powerful frameworks.

Flask: It is a micro framework of python that is used at back end to run webpages and integrate them to be ready to work with front end pages with all the integrated backend services.

SQLite DB: It is built-in database in Flask that is used to manage the user data and other system data.

Visual Studio Code: It is an open source IDE by Microsoft that provides workplace to work with above languages and frameworks.

Pandas: It is a Python data analysis library that is used to manipulate data in tables.

Beautiful Soup: It is a library written in python that is known for its ease of use, it is used for parsing of HTML and XML pages.

3.7 Database Requirements

Flask SQLITE DB is used for this system because of its built-in compatibility with python and its libraries. SQLITE is a simple, relational database that is best for this system, because this does not have any complex and large amount of user and system data to store.

3.8 Project Feasibility

Moving to a web scrapping technique from general roaming and searching for collecting data would affect the efficiency of gathering information in all manners. Many people would save their time because web scrapping would allow them to gather information at one defined place rather than searching randomly and gathering it and then making it at one place to make it suitable. Now the time has come to introduce this method of scrapping to achieve high level of efficiency. This project is the step-in right direction to lesser the frustration of common user who want to get relevant information at one place.

3.8.1 Technical Feasibility

Technically this project is more feasible because it provides more efficient way of gathering information from different sources. Contextual News Information Retrieval (CNIR) system is secure and less time consuming hence it provide more technical overheads than any other random system. The main module of the system is based on web scraping and python libraries, packages, frameworks, and architecture.

3.8.2 Operational Feasibility

This is an estimate of to what extent a developed system solves the problems that were identified earlier, and how well it takes the advantages of opportunities that were identified in the scope specification, to satisfy the requirements that were specified in requirements specification. The system is performing all its operations accurately. All functions include web scrapping, data gathering, displaying gathered data document.

3.8.3 Legal and Ethical Feasibility

Legal feasibility is basically the analyzation of at which extent your system is legally meeting the requirements that exist for implementation. The system is legally and ethically feasible as:

It does not violate any country law.

It is designed to fulfill people requirements of gathering data and minimize their effort on it to aid them in every possible manner.

Data of the user is completely secure, and it cannot be accessed without consent.

3.9 Summary

This section was about requirement specification, in which system requirements were discussed. System requirements include Interface, Functional, non-functional, database and resource requirements. Interface requirements include Software and Hardware requirements. FRs include must fill requirements whereas NFRs include common requirements of every system. At the end Project feasibility was discussed which include technical feasibility, operational feasibility and Legal & Ethical feasibility which means that the project is feasible within the time, budget and does not break or against the rules and regulations of the country.

CHAPTER 4

SYSTEM MODELLING

4.0 Introduction

In this section, the aid to get some type of models for graphic representation of the system from different type of perspectives of different stakeholders. The representation has been done by using UML.

Unified Modelling Language (UML) is a standardized modelling language which consists an integrated set of diagrams, that are used as an industry standard in the domain of Software Engineering [4]. In this chapter 4+1 view is used to represent the architecture of the system.

4.1 System Design

System Design is a procedure of theoretically specifying the development and usage of the product through architecture, models, and interactions, interfaces suitable data that is or can be processed by the system.

4.2 Design Approach

Since the system is subdivided into many subsystems therefore, Top-Down Design Approach has been used. By using this approach, the sub programs or modules of the system would also be identified and elaborated easily. The whole system is carried as a separate unit and partitioned into multiple sub systems and then each of these sub systems are carried as a separate unit and partitioned further. First defining the basic and generic model then going is depth and defining each part of it. The top-down approach for the system is given below.

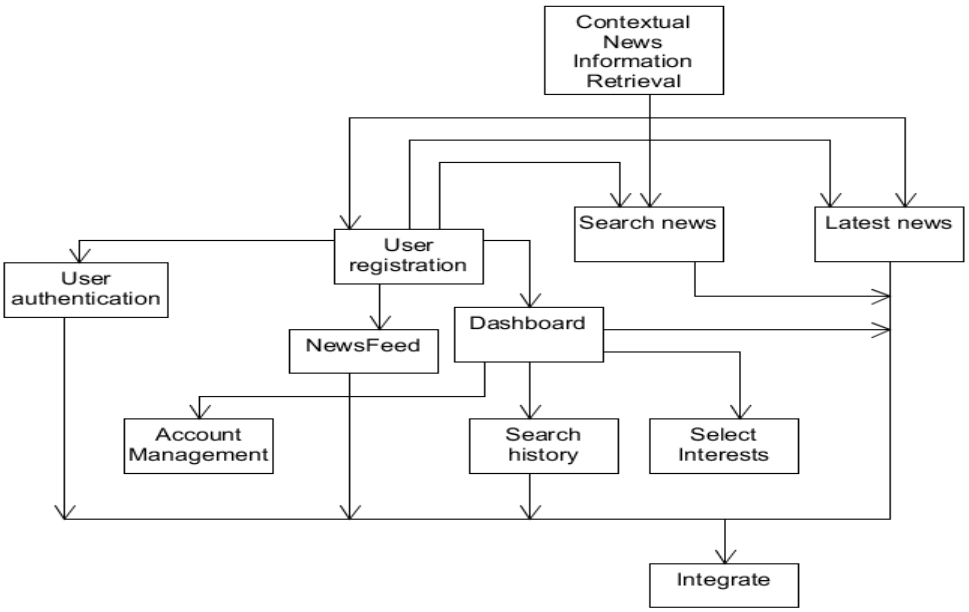

Figure 4.1 Top down approach for system

4.3 Interface Design

Interface design means the visual layout and placement of the system. An interactive or responsive manner layout used by the user of the system by some elements. These elements may be menus, buttons, radio buttons, check lists etc. The designs must not only be attractive but also display and explain the functionalities of the system.

4.3.1 High fidelity Prototype

High Fidelity Prototype is a high-level prototype which is very much attractive and shows the functionality, operating and workflow of the system. These types of prototypes are very close to the final design of the fully functional system. The level of detail and comprehensiveness in this type of prototype allows to test, analyse, and question the usability of the system workflow. High fidelity prototype allows you to examine usability questions in detail and make conclusions about the user behaviour.

4.3.1.1 Landing page

The following Figure 4.2 reflects the landing page of the system. The landing page includes navigation bar with search bar at the middle. The landing page also contains the logo of the system which is used by any website to advertise its product.

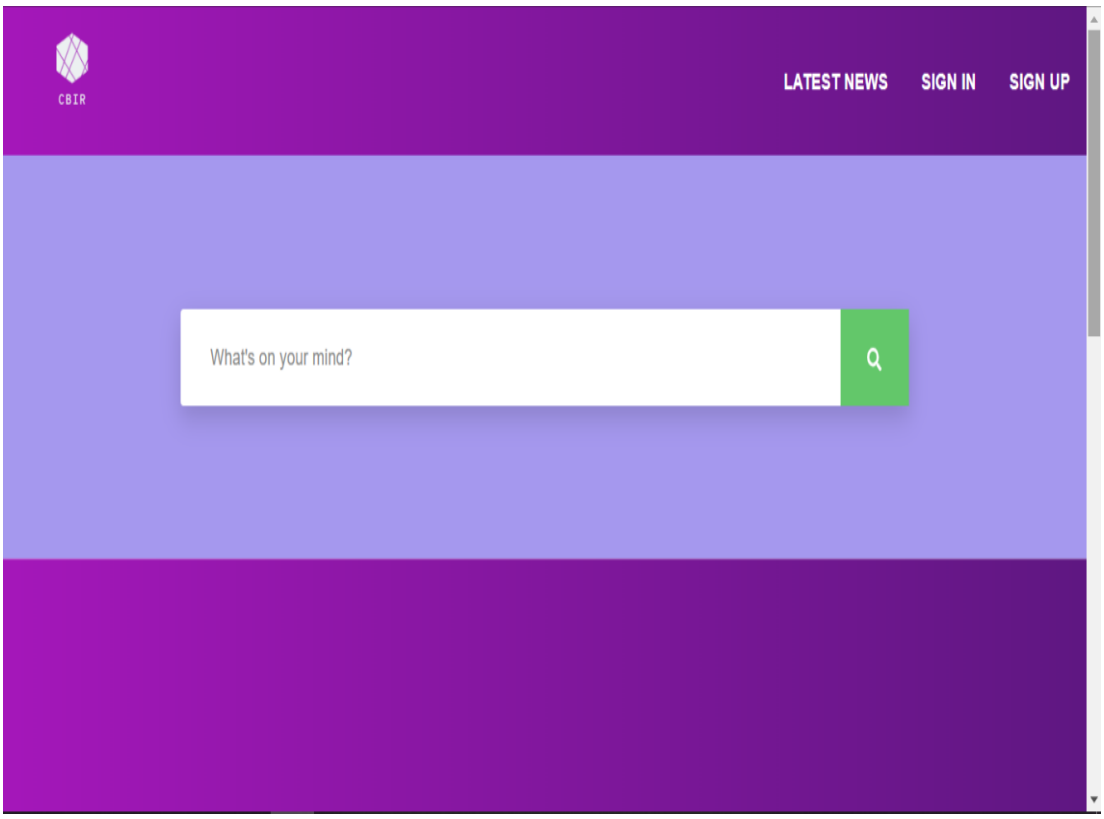

Figure 4.2 High fidelity view of landing page

4.3.1.2 Sign in

The following Figure 4.3 reflects the sign in page of the system. The sign in page includes email and password fields where user can enter its credentials.

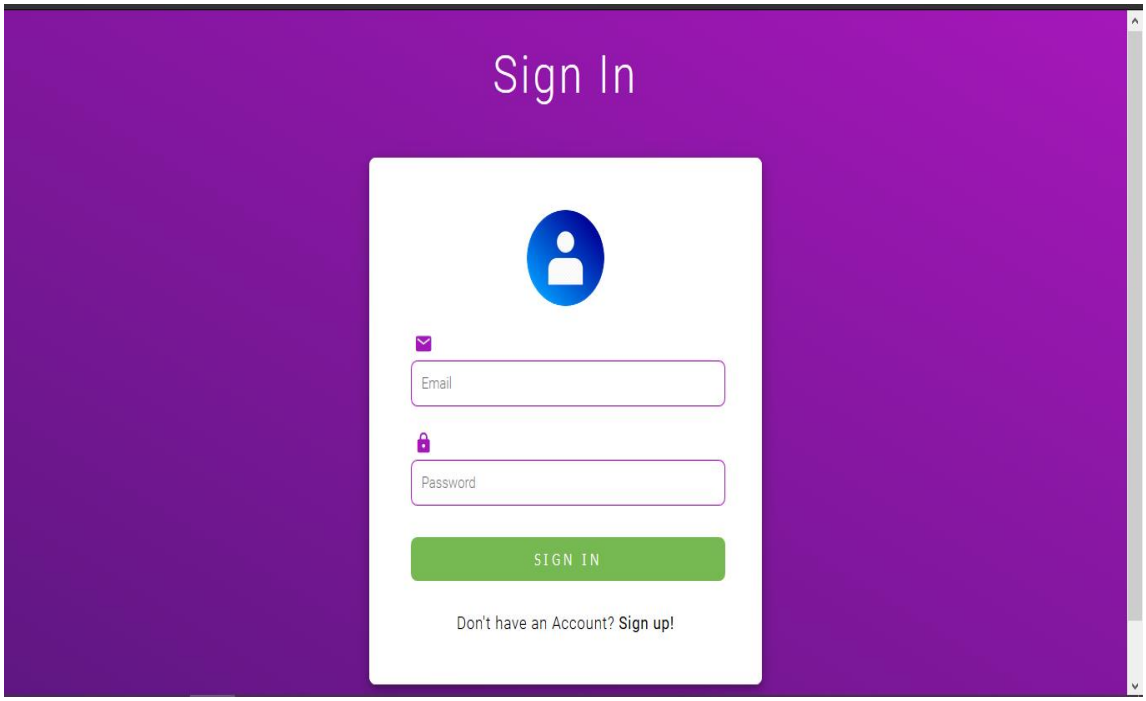

Figure 4.3 High fidelity view of sign in page

4.3.1.3 Sign up

The following Figure 4.4 reflects the sign-up page of the system. The sign up page includes name, email and password fields where user can enter its details to get registered.

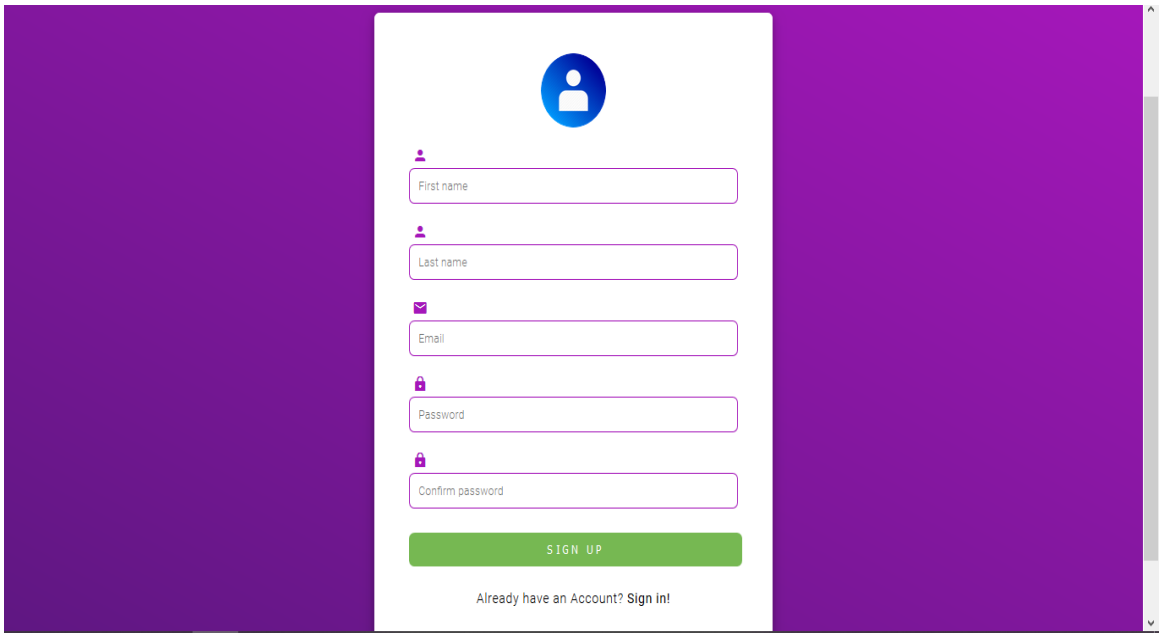

Figure 4.4 High fidelity view of sign up page

4.3.1.4 Search news

The following Figure 4.5 reflects the search news information page of the system. The search news page shows news against the user query.

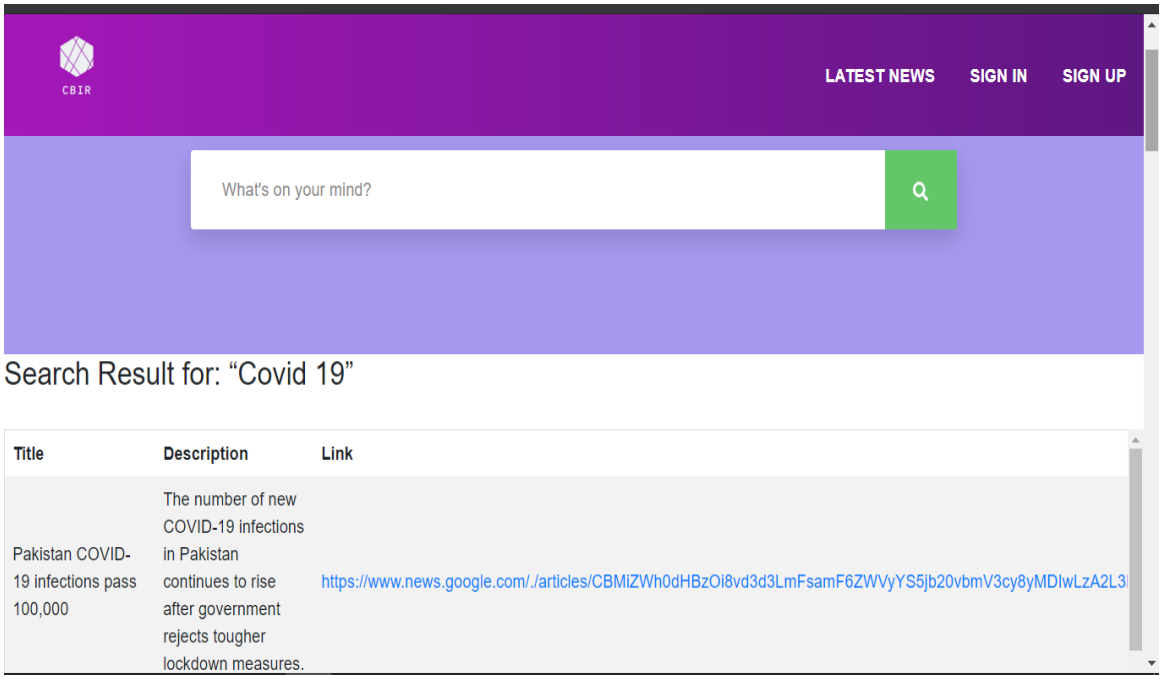

Figure 4.5 High fidelity view of search news

4.3.1.5 Homepage for registered user

The following Figure 4.6 reflects homepage of the system for registered user. The homepage shows enrich feature content to registered user containing latest news and newsfeed.

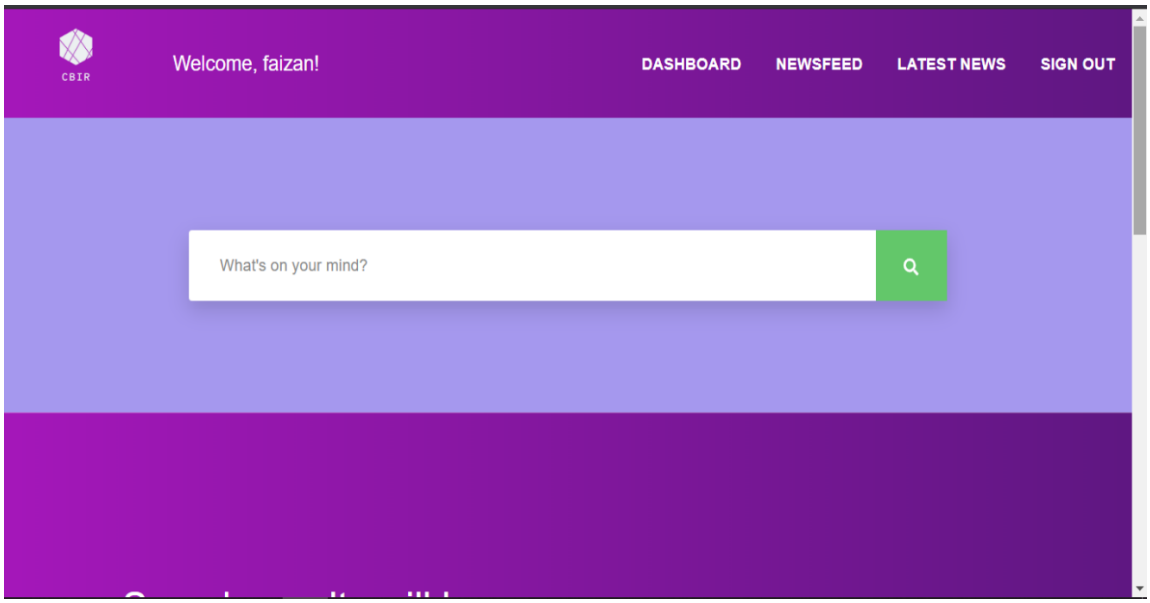

Figure 4.6 High fidelity view of registered user home page

4.3.1.6 Latest news

The following Figure 4.7 reflects the latest news page of the system. The latest news page shows latest news to user from five news websites of Pakistan.

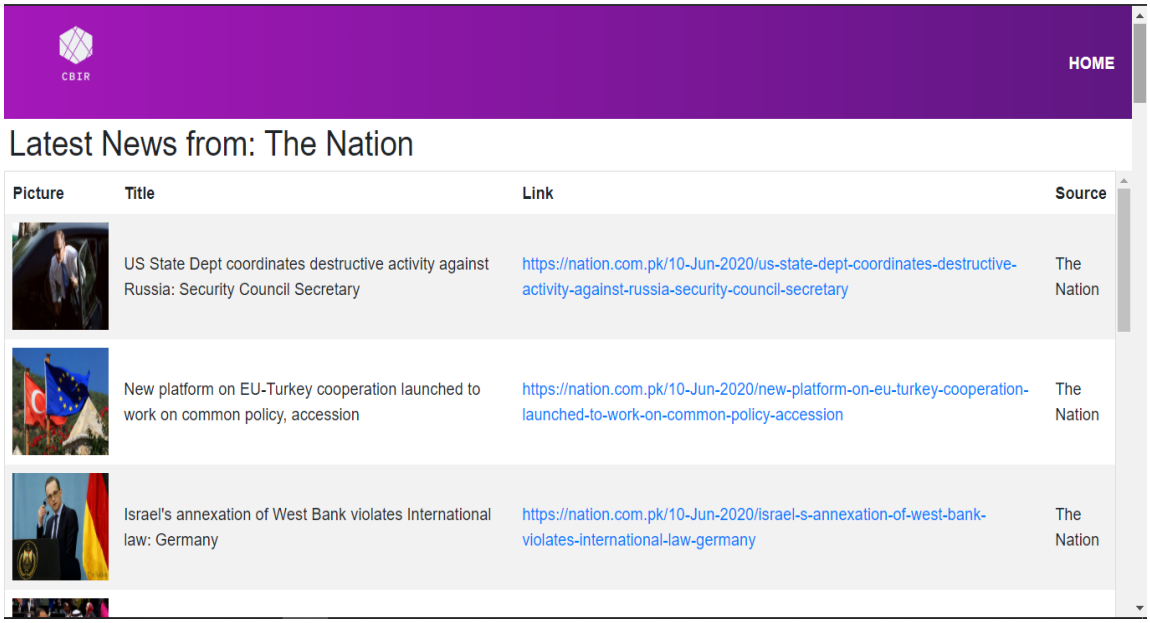

Figure 4.7 High fidelity view of latest news

4.3.1.7 Newsfeed

The following Figure 4.8 reflects the newsfeed page of the system. The newsfeed page shows news related to news interests selected by user.

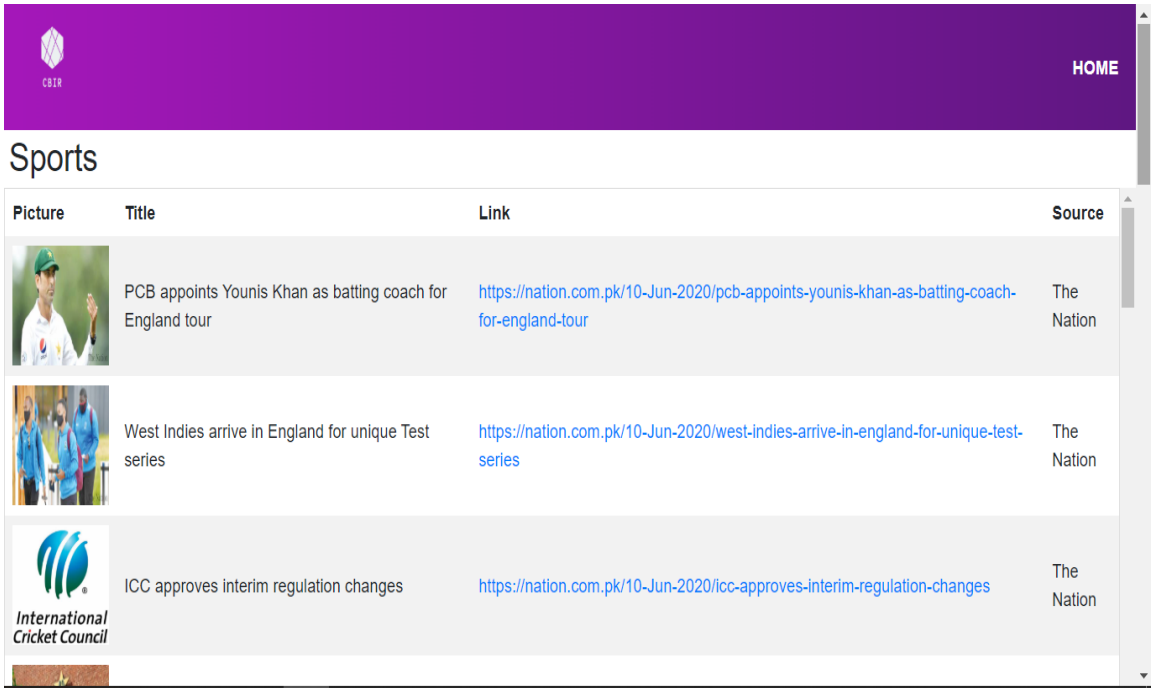

Figure 4.8 High fidelity view of newsfeed

4.3.1.8 Change account name

The following Figure 4.9 reflects the change account name page of the system. Change account name contains fields for first name and last name of the user.

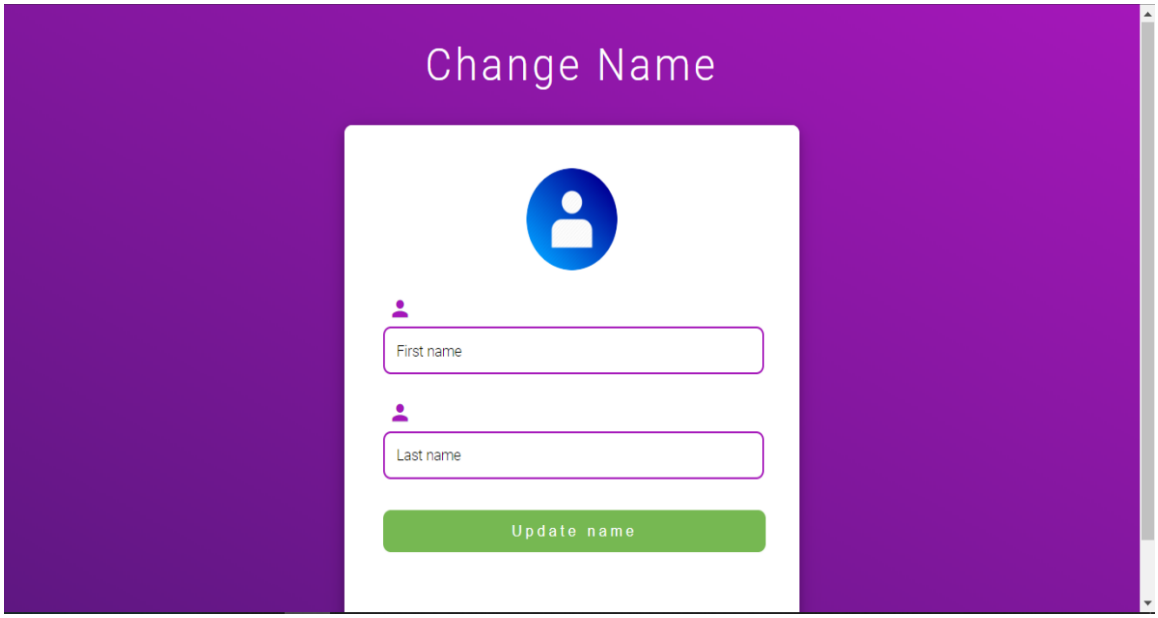

Figure 4.9 High fidelity view of change account name

4.3.1.9 Change account password

The following Figure 4.10 reflects the change account password page of the system. Change account password contains fields for old password and new password for user.

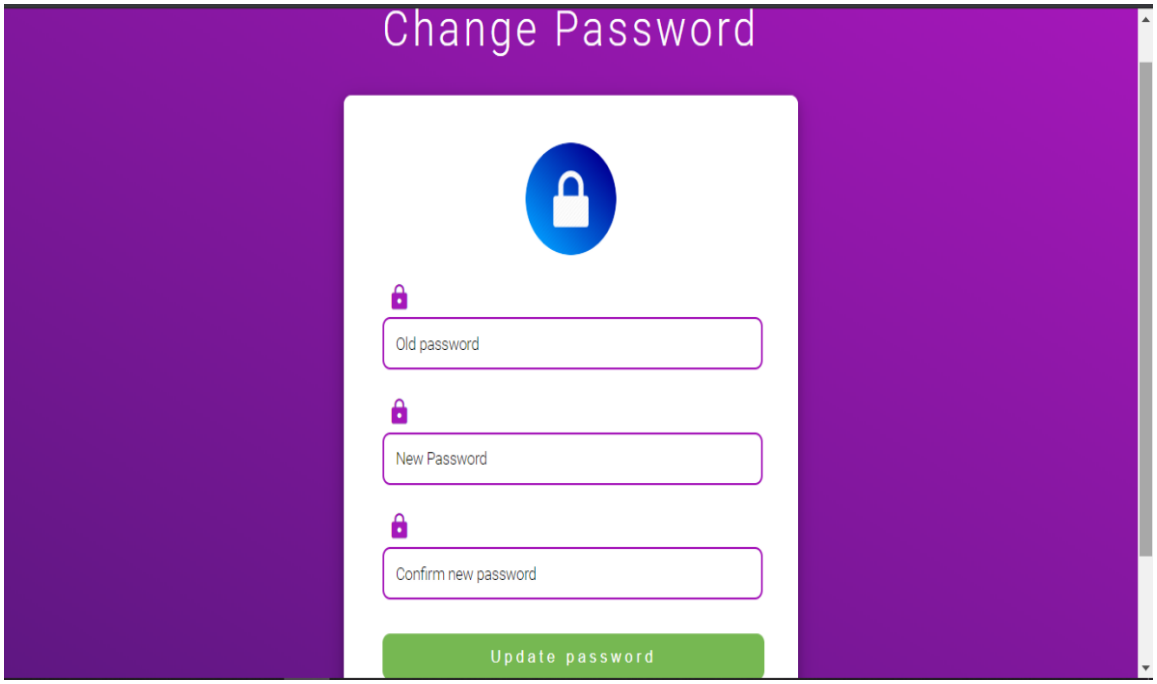

Figure 4.10 High fidelity view of change account password

4.3.1.10 Dashboard

The following Figure 4.11 reflects the dashboard page of the system. The dashboard page shows the side navigation panel and news search history at the middle of page.

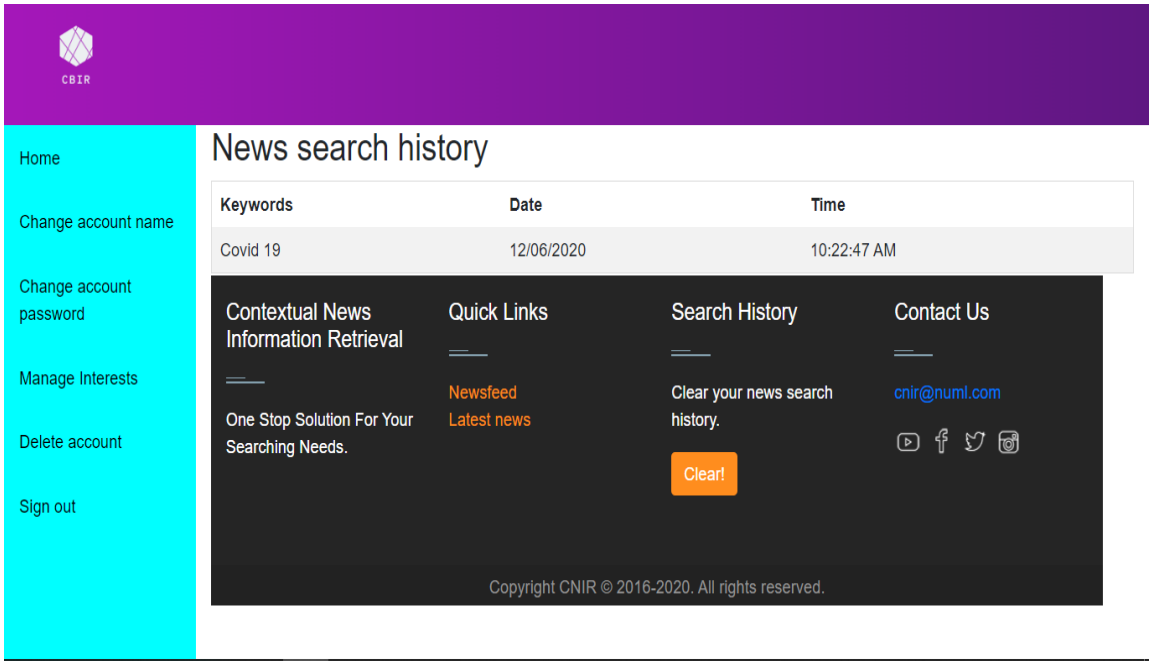

Figure 4.11 High fidelity view of the user dashboard

4.3.1.11 Select news interests

The following Figure 4.12 reflects the select news interests of the system. The select news interest page shows a dialog containing all the news interest for users to choose from.

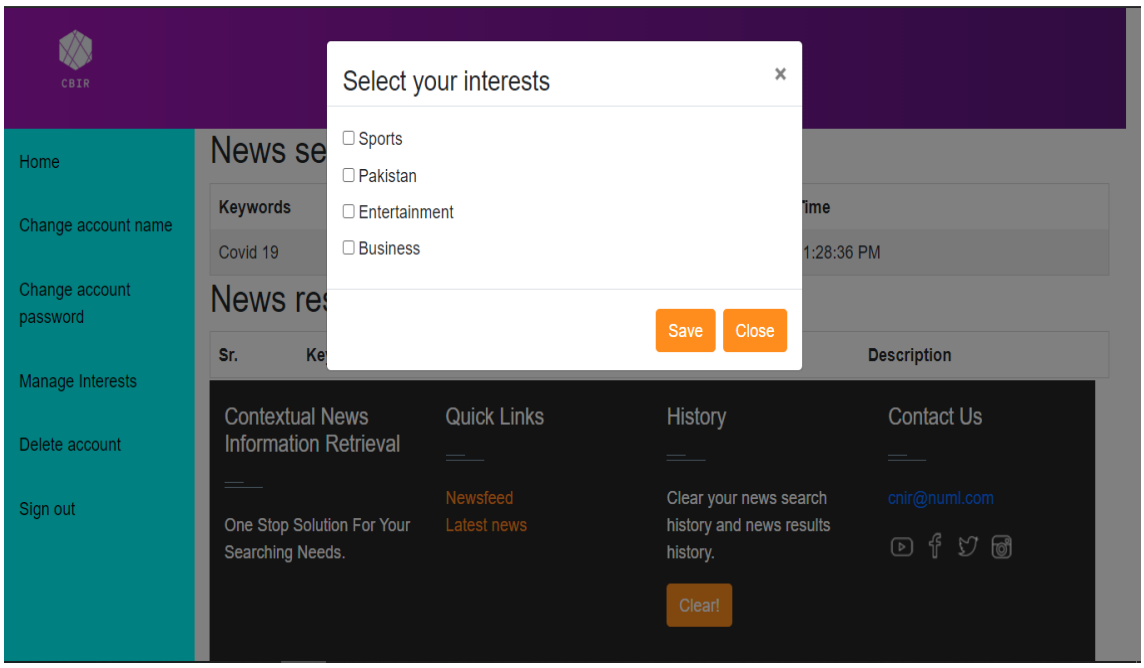

Figure 4.12 High fidelity view of select news interest

4.3.1.12 Search history

The following Figure 4.13 reflects the search history page of the system. The search history is displayed in the dashboard page with a button to clear the search history of the users.

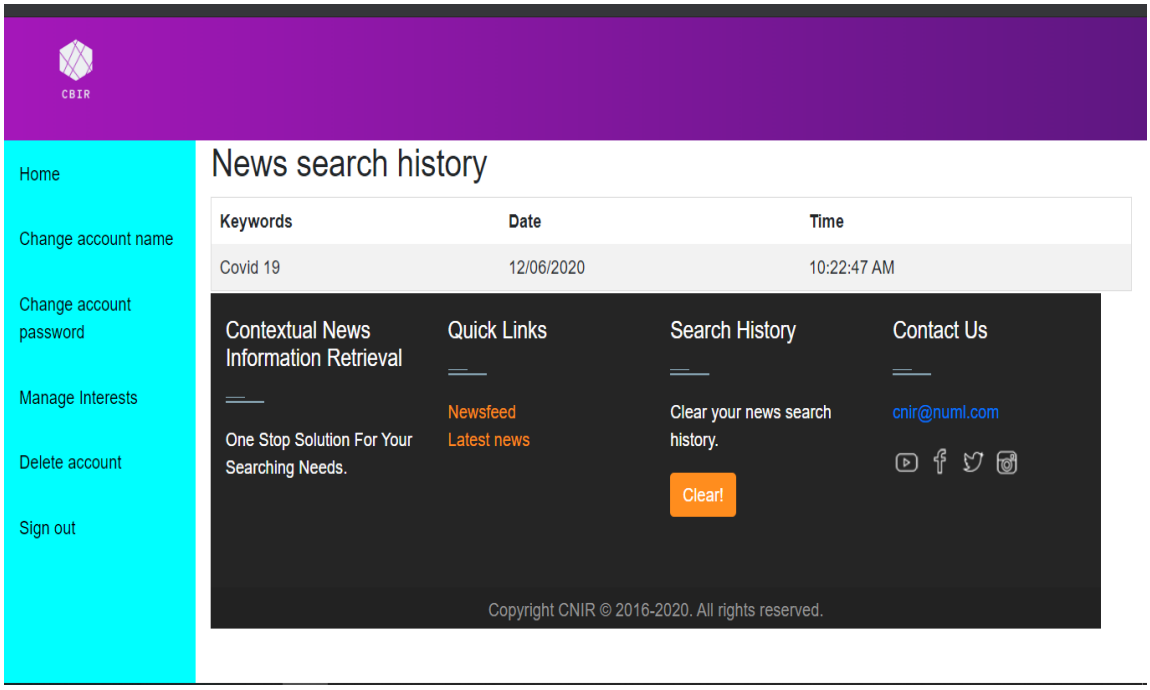

Figure 4.13 High fidelity view of search history

4.4 4+1 view Model of Architecture

All views which are application to this project are logical view which includes class diagram, and ERD of the system. Process view which includes activity diagram, state machine diagram and sequence diagram. Development view which includes component diagram and physical view which include deployment diagram. All of them are described below:

4.4.1 Logical View

Logical view demonstrates the logical parts of the system and contains the information about these parts. This view helps in determining the logical connectivity of different parts of the systems.

4.4.1.1 Class diagram

It is a structural diagram in UML which demonstrates about the stagnant structure of the system. This diagram consists of classes, attributes, methods, and relationship among different objects that how the objects are related to each other. The class diagram of the system is displayed below in Figure 4.14.

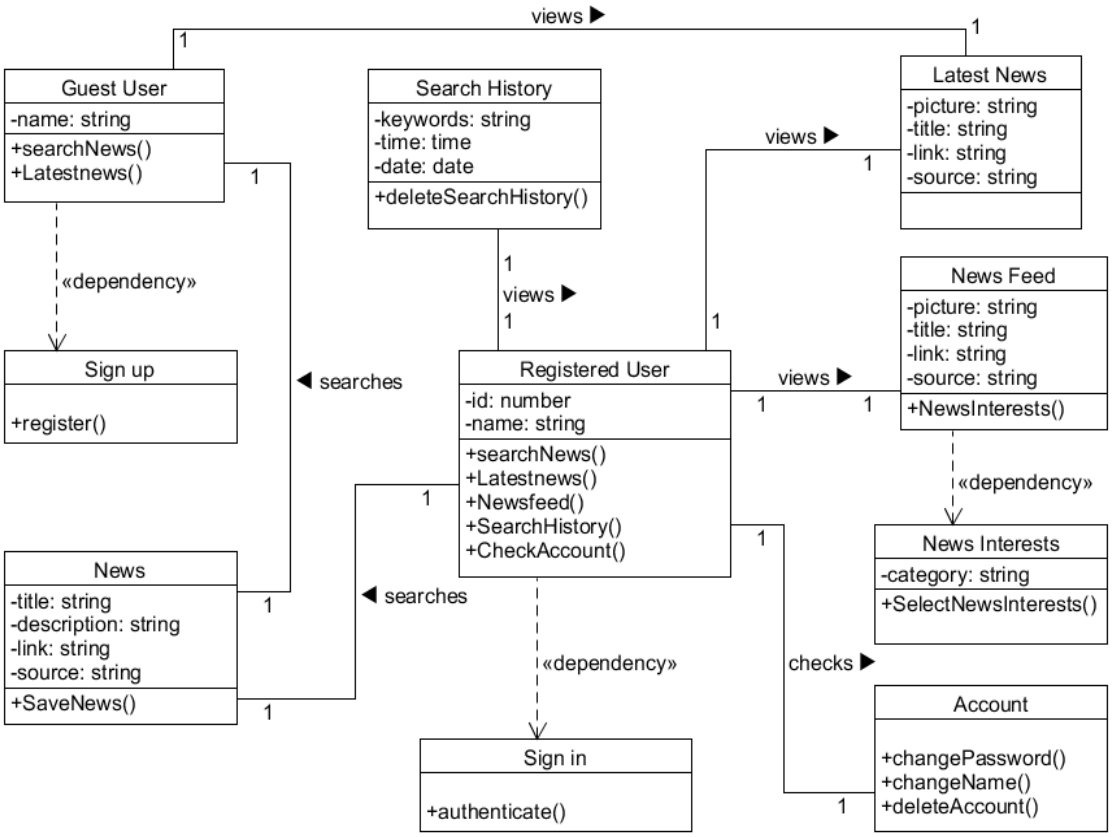

Figure 4.14 Class diagram of the system

4.4.1.2 ERD

Entity Relationship Diagram (ERD) of the system is given below in Figure 4.15.

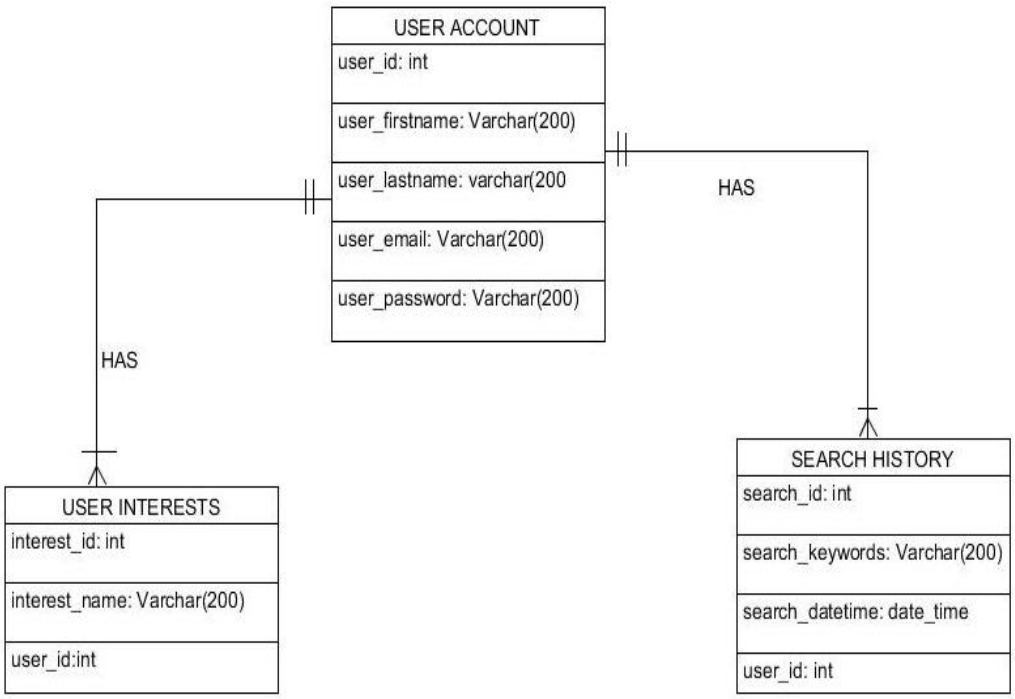

Figure 4.15 ERD of the system

4.4.2 Process View

Process view demonstrates workflow and moving of functionality of the system. It tells about concurrent processes of the system. This view helps in analysing the performance and availability of the system

4.4.2.1 Activity diagram

Activity diagram is a behavioural diagram in the UML. Figure 4.16 portrays the dynamic facets of the system below.

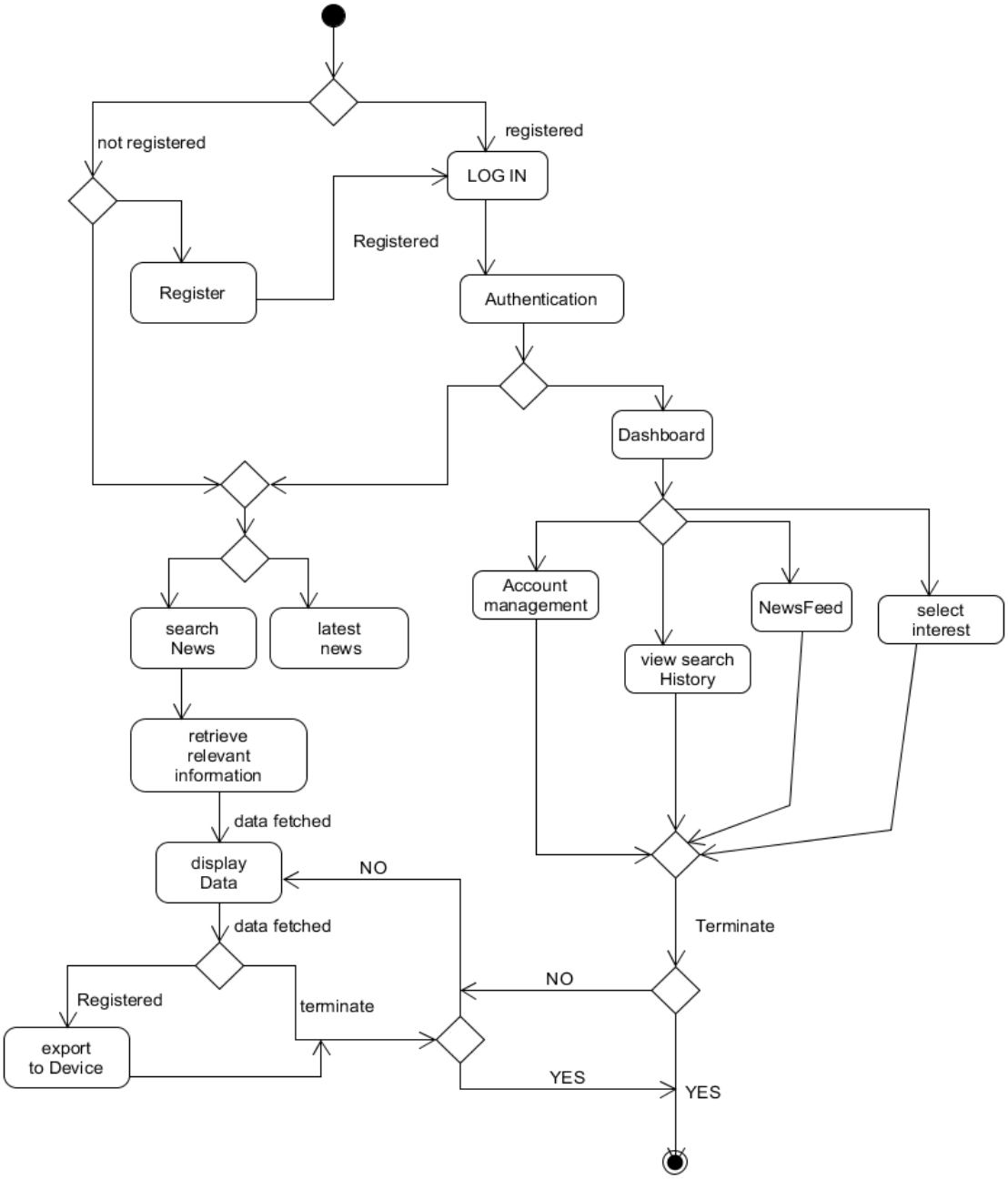

Figure 4.16 Activity diagram of the system

4.4.2.2 State machine diagram

State machine diagram demonstrates discrete actions of the system. Figure 4.17 shows entities, states of the system and connect flow between them.

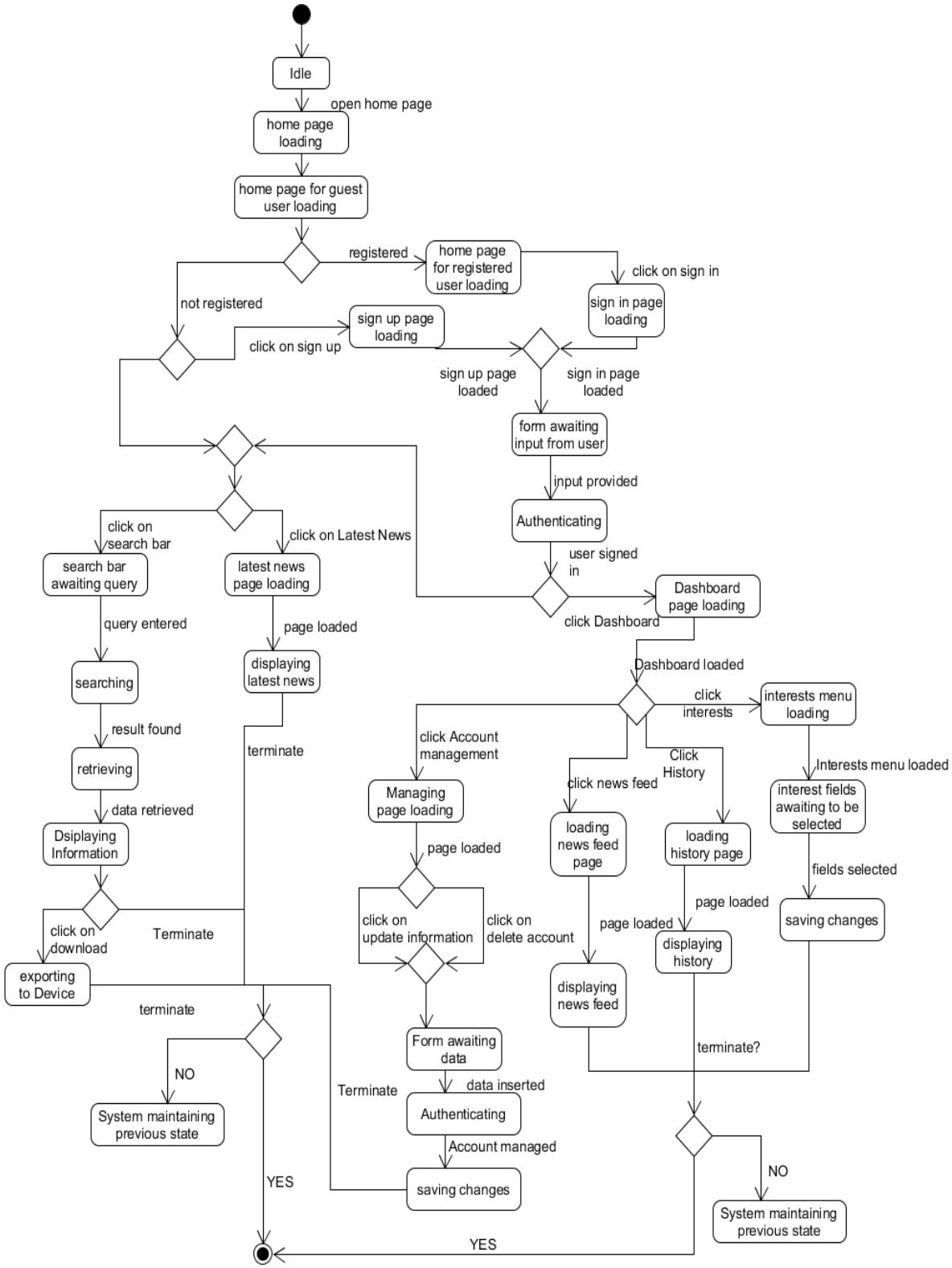

Figure 4.17 State machine diagram of the system

4.4.2.3 Sequence diagram

It helps to signify the behaviour of the system in response to the communications and involvement of the user. Functionality of each module is demonstrated by providing dynamic results based on the input of the user. It is called sequence diagram because it shows the behaviour of the system and its reply to the users interactions in a sequential manner. It is also called event diagrams because it describe interactions among classes in terms of an exchange of messages over time.

4.4.2.3.1 Sign In

The sequence diagram for sign in is given below in Figure 4.1. The user is signing in by entering email and password. In response, the system is validating the user credentials and displaying the message.

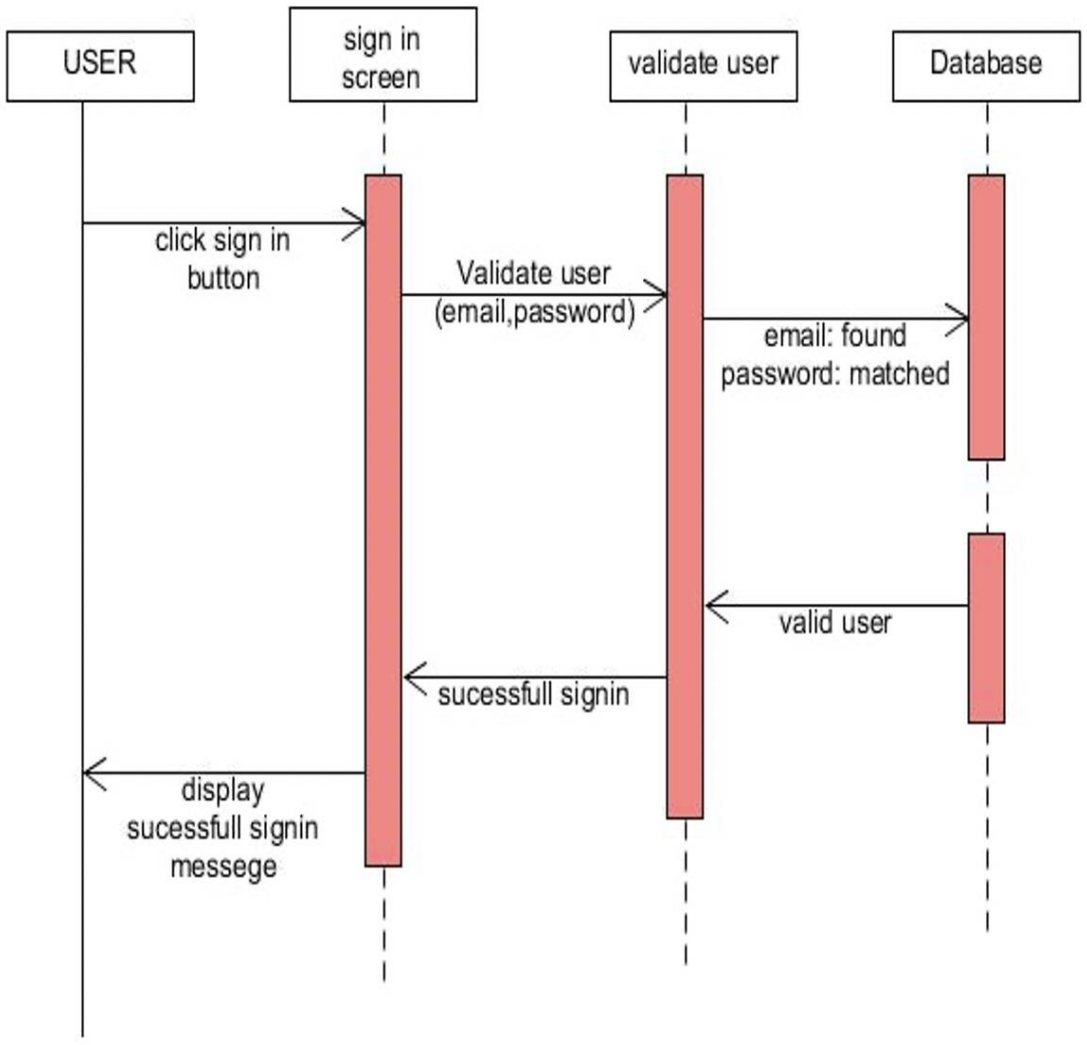

Figure 4.18 Sequence diagram for sign in

4.4.2.3.2 Sign up

The sequence diagram for sign up is given below in Figure 4.19. The user is entering sign up details. In response, the system is registering the user and displaying message.

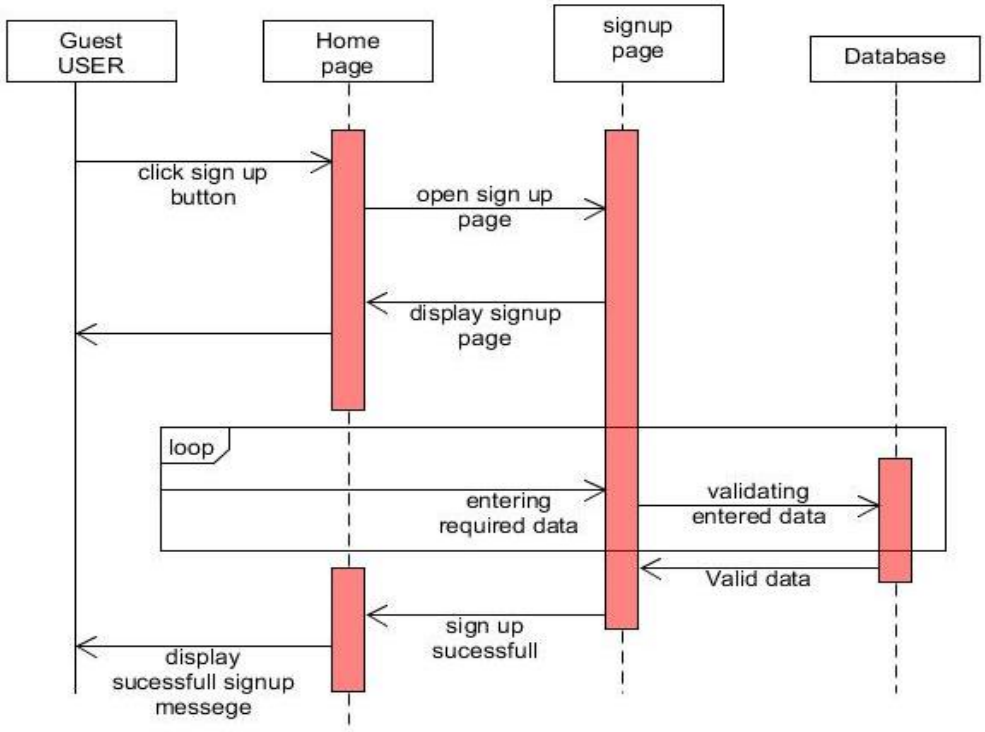

Figure 4.19 Sequence diagram for sign up

4.4.2.3.3 Search News

The sequence diagram for search news is given below in Figure 4.20. The user searching the news. In response system is querying the news information.

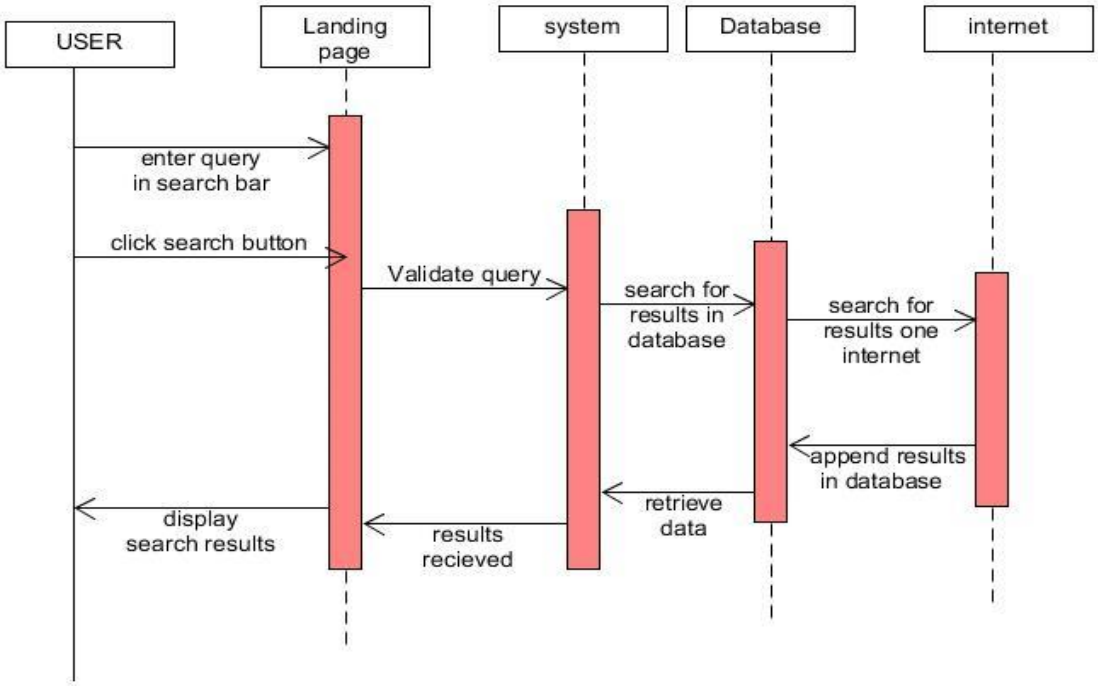

Figure 4.20 Sequence diagram for search news

4.4.2.3.4 Search history

The sequence diagram for search history is given below in Figure 4.21. The user is opening dashboard page. In response the system is showing search history on dashboard page.

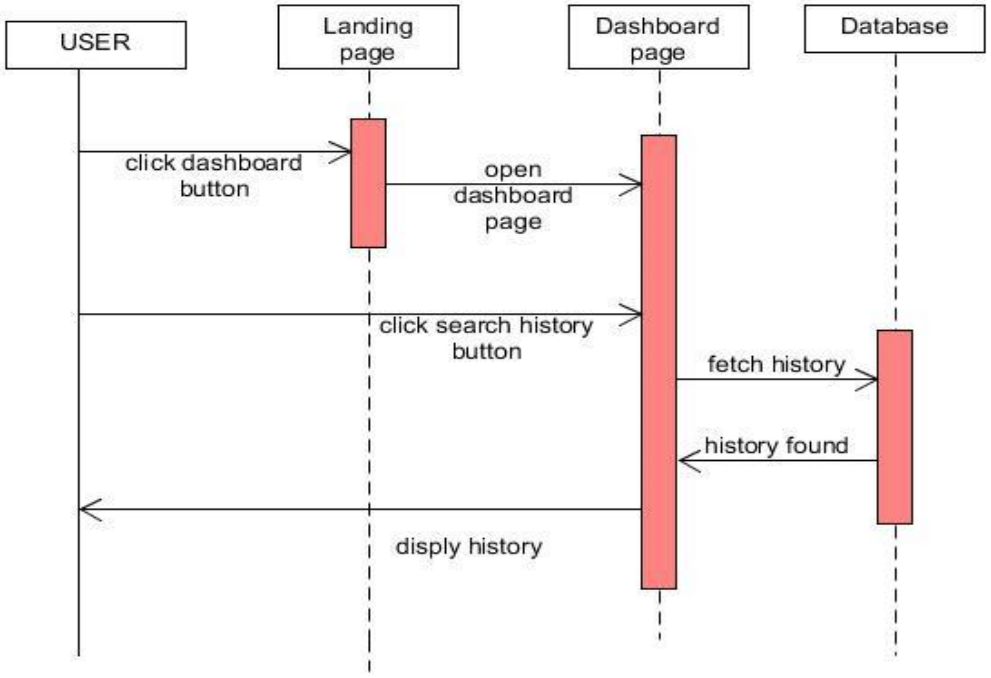

Figure 4.21 Sequence diagram for search history

4.4.2.3.5 Newsfeed

The sequence diagram for newsfeed is given below in Figure 4.22. The user is opening newsfeed page and in response system is showing news related to interest of user.

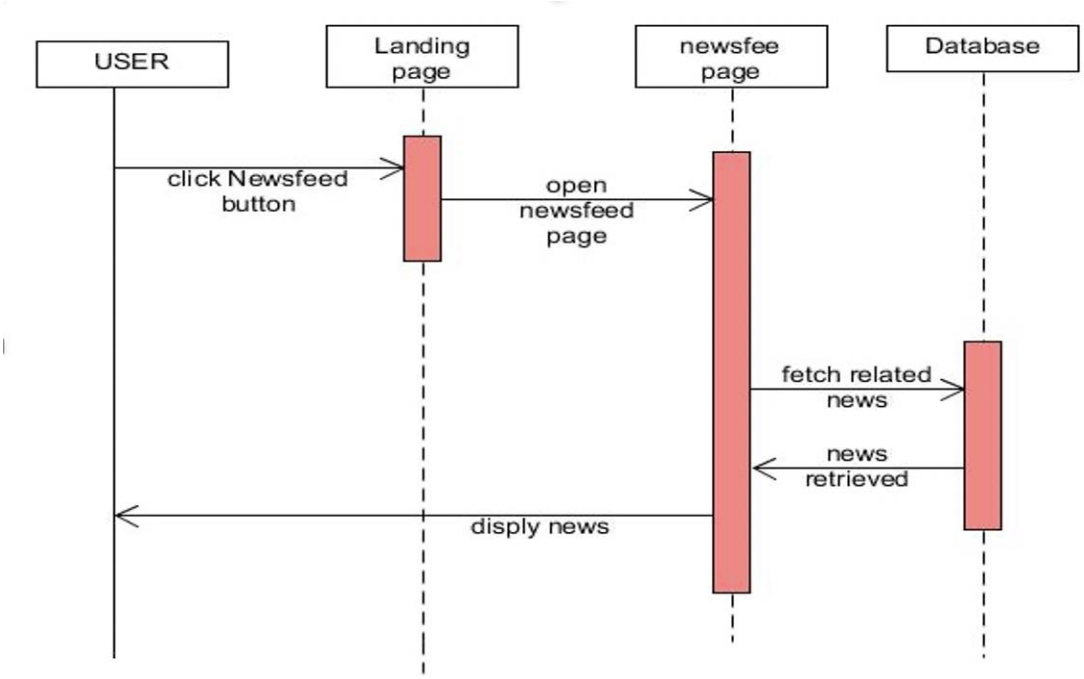

Figure 4.22 Sequence diagram for Newsfeed

4.4.2.3.6 Latest news

The sequence diagram for latest news is given below in Figure 4.23. The user is opening latest news page and in response system is showing real time news to the user.

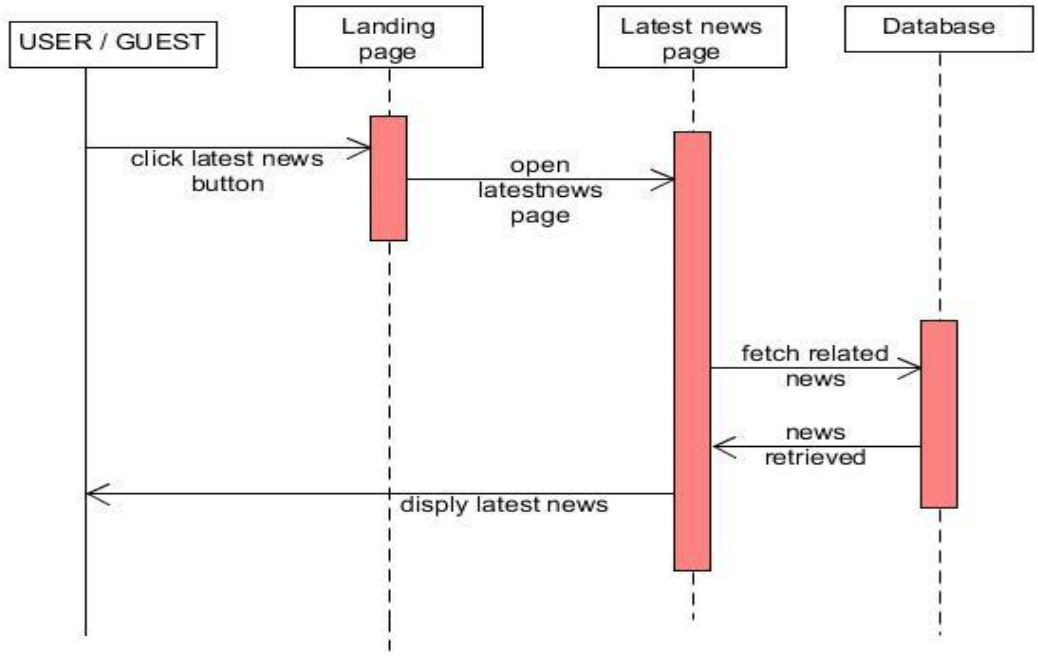

Figure 4.23 Sequence diagram for latest news

4.4.2.3.7 Select news interests

The sequence diagram for select news interests is given below in Figure 4.24. The user is selecting news interests and in response system is saving news interests selected by user.

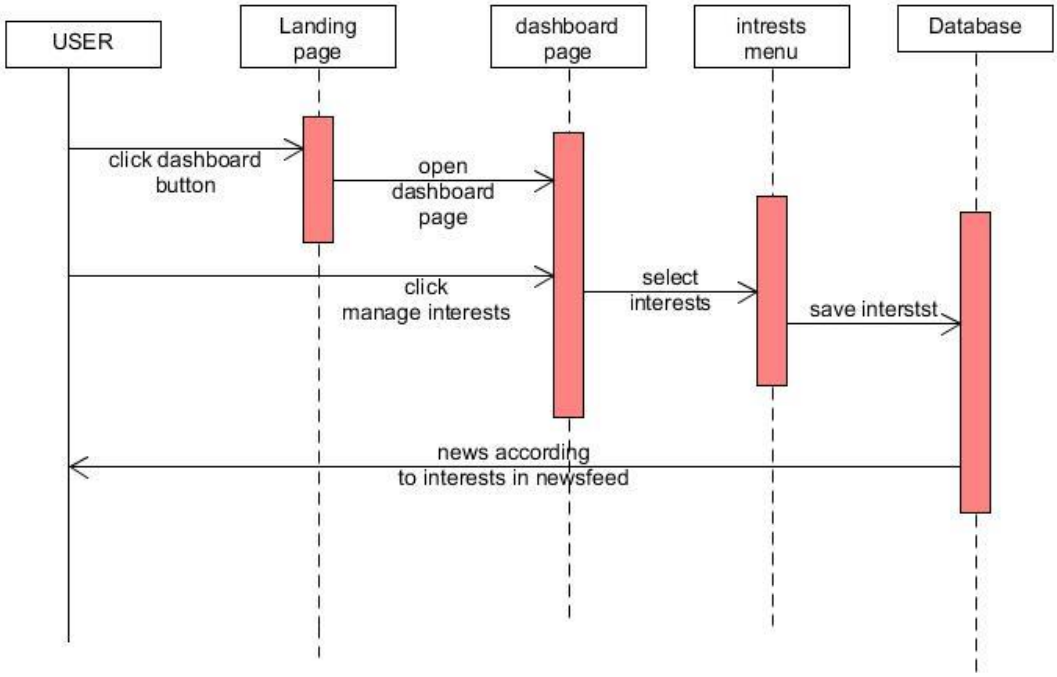

Figure 4.24 Sequence diagram for select news interests

4.4.2.3.8 Save news

The sequence diagram for save news is given below in Figure 4.25. The user is downloading news and in response system is saving a file containing news into local machine.

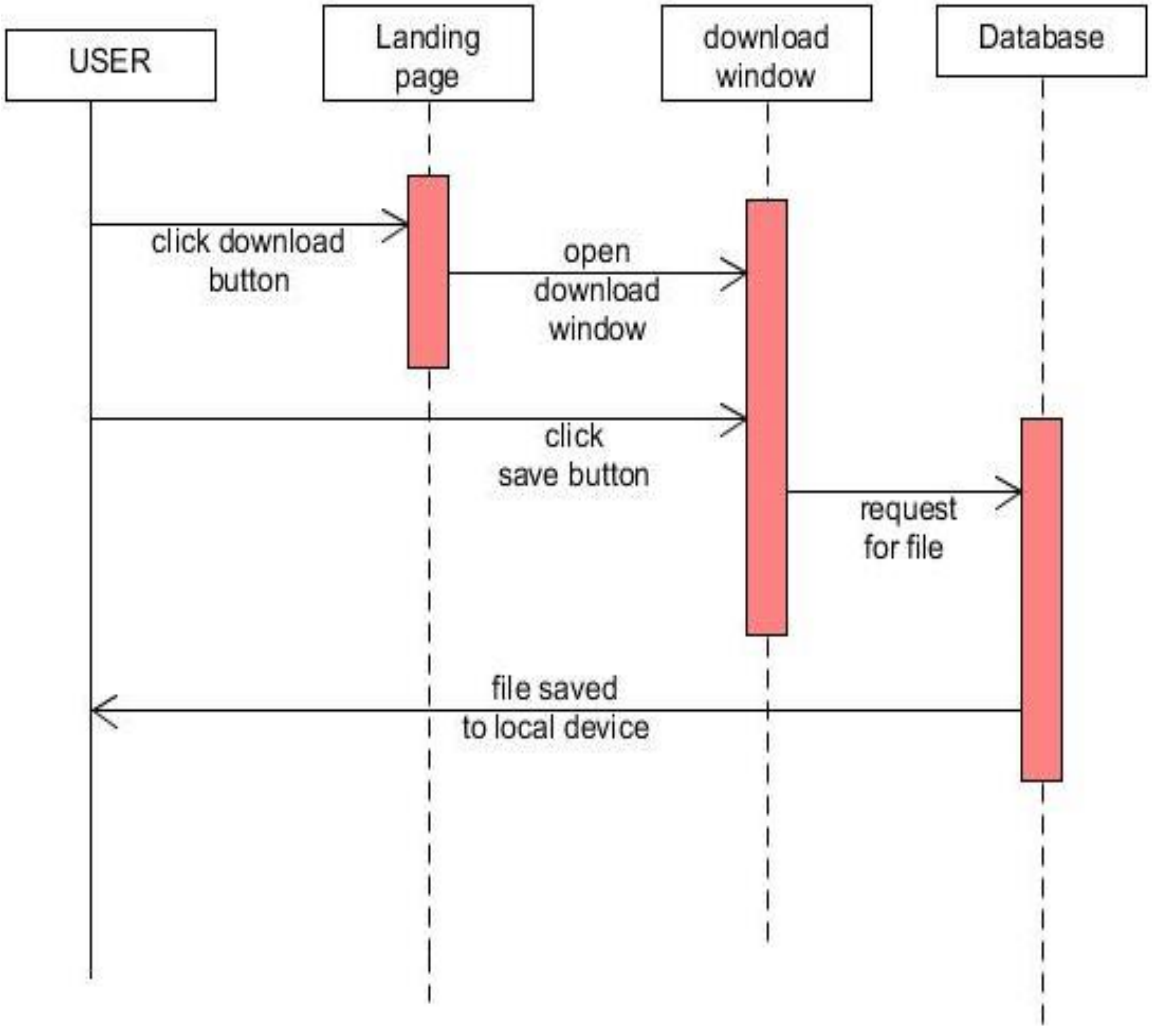

Figure 4.25 Sequence diagram for save news

4.4.3 Development View

This view helps to model and demonstrates the components, modules, and subsystems of the main system. This view is also known as the implementation view. It uses the UML Component diagram to describe system components.

4.4.3.1 Component diagram

Component diagrams are used in capturing the physical parts of Object-Oriented Programming (OOP) based applications that are used for capturing, identifying, and specifying component-based aspects. It does not describe the functionality of the system but it describes the components used to make those functionalities.

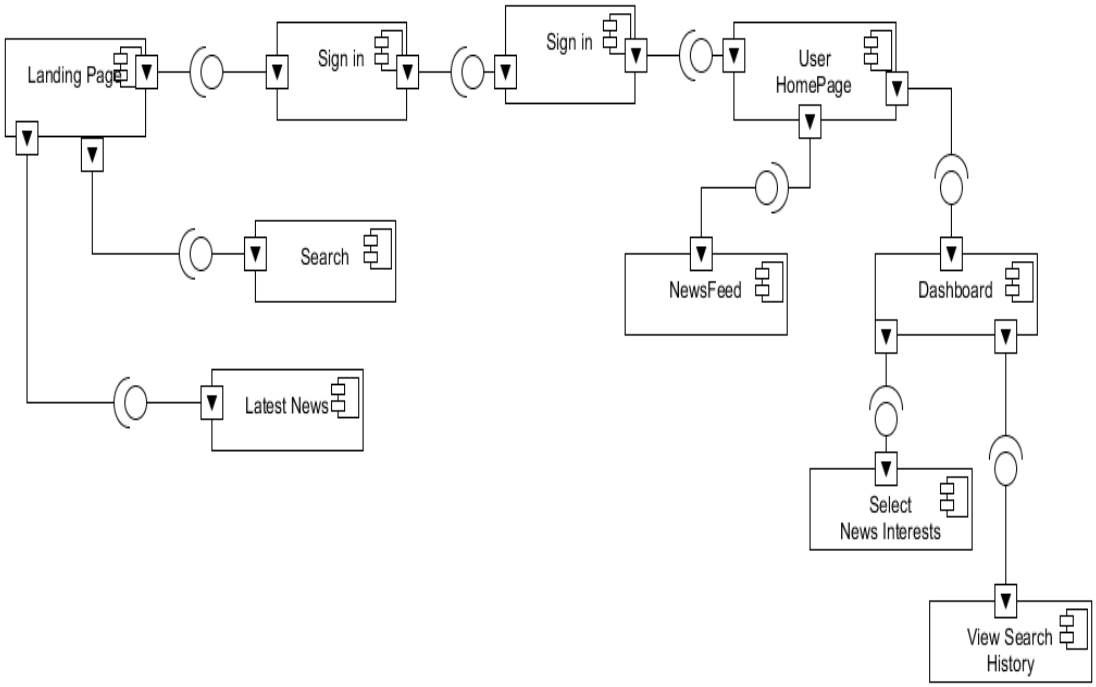

Figure 4.26 Component diagram of the system

4.4.4 Physical View

This type of view tells the physical and structural deployment of the system and the type of network and physical layout this system is using and is designed. Deployment diagram

This figure explains physical deployment of application under development. The arrangement of nodes, what type of functionality is performed by the nodes. It helps to figure out the stability and availability of application.

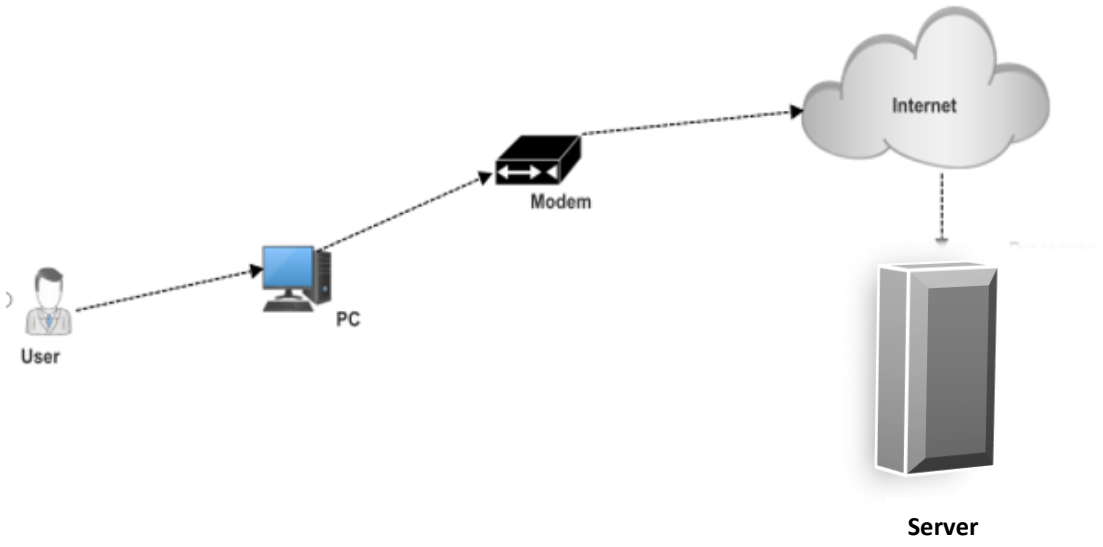

Figure 4.27 Deployment diagram of the system

4.5 Summary

This chapter elaborates the complete modelling of this application from start to end. A high-level representation of the system has been given.

In the activity diagram, the dynamic behaviour of the web-based application and all the alternates of different activities are represented. In sequence diagrams, the different responses of the system opposing to different interactions of the user and the sequence of steps required to perform the functionality are discussed. Component diagram aids in representing the various subsystems and modules of this application to easily enhance the functionality.

CHAPTER 5

IMPLEMENTATION

5.0 Introduction

In this chapter, the main implementation of the components of system are discussed. Module by module implementation of the application is explained. How each module was developed, and which library, framework, or service was used. Detail of implementation of every module of the system is discussed. The working of the library, framework, or service that was used in the module is also explained in detail.

5.1 Modules of Project

Modules of the system are given below with their implementation details.

5.1.1 User Registration

The module user registration allows the guest users to register (sign up) themselves in the system database, the user enters the first name, last name, email, password to get registered on the system.

The module user registration allows registered users of the system to authenticate (sign in) themselves into the system, the system prompts the email and password to authenticate on the system.

5.1.2 News Search

The module news search of the system enables anyone to enter something and search the news. User can enter the keyword related to the story, and then website scraps the required data, which is the best match for these keywords. The user can only read the information on-site and visit the site via the link which is provided with the news.

5.1.3 Save News

One of the most important modules of this system is save news. This module provides a platform to record data extracted from relevant sources. Users can use this platform to have factual data in the form of document and download it in their local machine without wasting time and effort.

5.1.4 Search History

This Module of the system allows users to check their previous searches. The search history also increases the searching capability of the system. The logged in user goes into dashboard where the history option is provided. The system maintains the complete details of the user history in that section, and the user can clear the history.

5.1.5 Latest News

This Module of the system allows users to see latest news from five Pakistani English news websites. Both guest user and registered users can click on latest news on their respective homepages to see the latest news from those five Pakistani English news websites which includes The Nation, Dawn, Pakistan Today, Daily Pakistan and Express Tribune.

5.1.6 News Interests

This Module of the system allows to select news interests from four categories. The logged in user clicks the dashboard where the “Manage Interests” is provided. The system maintains the news interests selected by user in that section, and then section of newsfeed updates accordingly.

5.1.7 Newsfeed

This page allows users to see newsfeed based on the news interests that the user had selected. The user logs in to the system to open homepage where the user clicks on newsfeed to see the newsfeed based on the selected interests.

5.2 Tools and Technology

The following are tools, libraries, framework, and services that were used in developing the modules of the system. These are flask, pandas, beautifulsoup, tesseract-ocr, requests, numpy, and sklearn.

5.2.1 Flask

Flask is a framework or micro-framework which is supported by python programming language, and it is mostly written in python. It is maintained by community of individual python developers and companies. Flask can be used as a base in the development of different functions or applications, as it is optimal for fetching changing data that needs to be recorded. Flask is for developing the modules of the system.

5.2.2 Tesseract-OCR

Tesseract OCR is an open source library written in python that helps recognize characters in images pdfs etc, it supports more than 100 languages. Other languages which are not included can also be extracted by training it. It is optimized for complex character recognition and simple character recognition. Tesseract is licensed by Apache and maintain by Google from 2006 [5].

5.2.3 Pandas

Pandas is a free to use open source library written in python programming language, it is known for its easy to use methods of data analysis and handling CSVs, its most common use is data manipulation [6]. Pandas developers developed it for different mathematical and time series analysis, including different tabular data. Pandas data frames have been used to save data to a CSV file in this project, as pandas makes working with csv files a lot easier. Consequently, Pandas is a free library which is used for different useful functionalities.

5.2.4 Beautiful Soup

BS4 is another open source library based on python 3.8 and is used for parsing. Parsing is done for HTML and XML documents, it provides many vernacular ways to navigate through the parse tree, parse trees can be modified and searched easily by using this. It has been recognized as a major time saver for programmers [7]. In the project, BS4 has mainly been used to navigate through the target page's html and get the required information.

5.2.5 Requests

Requests is a library developed for the better understanding of extracting data from different HTML and XML, and it is human friendly and easy to use. HTTP requests can be easily sent using Requests library, there is no manual query adding work needed in URLs by using this [8].

5.2.6 NumPy

NumPy is an important library of python programming language mainly designed for dealing with matrices and to perform numeric operations as lists in python doesn't allow any numeric operations; matrices involve two dimensional, three dimensional, and so on and other mathematical programming calculations. It is also popular in dealing with different arrays. In this project, NumPy Arrays are not directly required, instead the information is stored in Pandas data frames, but Pandas is a library that is built over NumPy, so it can be say that NumPy is a pre-requisite for Pandas to work on any system, same is the case with SKLearn.

5.2.7 SKLearn TFIDF Vectorizer

SKLearn is one of the most known libraries and it provides many supervised and unsupervised learning algorithms. It is based on some common technologies that are already known. These include numpy, pandas, and matplot. Some common functionalities

that SKlearn provides include, clustering, pre-processing including min-max normalization etc [9]. SKLearn TFIDF vectorizer is mainly used to tokenize the document, analyse the text for reverse document frequencies and encode the material, and it is also used for weighting the document and learning the vocabulary of it. In this project it has been used to find term frequency and inverse document frequency for each word of the entered query so the system can get the context of what the user is searching.

5.2.8 SKLearn Cosine Similarity

Scikit Learn Cosine Similarity is used to estimate the similarity in documents irrespective of their sizes, it's basically a mathematical model that measures the cosine of angles between vectors [10]. SKLearn Cosine Similarity is used to find the similarity between the documents, it mainly calculates its frequency of similarity. In this project, when a query is entered by a user, the system first tries to find that news in the data that has already been retrieved and stored, to save the time. For this purpose, it is important to determine that whether the news that is retrieving matches the query of the user or not. So, the Cosine Similarity Method is used to determine the similarity of context as it ignores common words (is, am, are, he/she) and finds the similarity based on important words. The entered query is first matched with already retrieved data and if the similarity is above 50% then top 5 news are shown to user.

5.3 Summary

This chapter focuses on the module implementation of the system. Tools that were used in developing were discussed briefly. Flask, Pandas, NumPy, Beautiful Soup and Requests libraries are used in the back end for the development of the modules of the system. Flask was used as the main framework for integration of backend and frontend components. NumPy is a prerequisite for Pandas, that is why it is used, Pandas was used to save news information to data frame and then save it in a file to maximize convenience and minimize resource utilization. SKLearn is library for Natural Language Progressing. Scikit Learn or SKLearn is a very popular library written in python for Machine Learning algorithms, its known as one of the best libraries to work with NumPy arrays as it supports many scientific and numeric libraries. SKLearn TFIDF Vectorizer was used to tokenize the query in separate words and determine the inverse document frequency of different words to know about the important terms of the document or query. SKLearn Cosine Similarity is used to compare two documents and find the similarity index between them, in this project user's

entered query is compared with instances of different news residing in csv file, to get the similarity, the news which get their similarity index more than 50% are shown. HTML, CSS, JavaScript, and bootstrap acts as a foundation for developing front end design, validation with modular architecture and applications using plug-and-play components. Detailed discussion about the implementation of different modules of the application is discussed.

CHAPTER 6

TESTING, ANALYSIS AND VALIDATION

6.0 Introduction

All the necessary reporting and documentation required for testing and validation in this chapter. As it is known that in testing, the project is run with the intention to find the errors and remove them to improve the functionalities of the system. The test cases are utilized to test all the functionalities of system by analysing and validating the results gathered from those test cases. As it is known that testing is a process of executing the system in real time and then analysing and evaluating the results. The goal is to find and correct as much errors and defects before launching the system in the market. Some of the testing techniques that have been used in the testing process are unit testing in which a specific capability or a functionality of system is tested, a part from that have been used integration testing which is performed after some chunks of the whole system are integrated to test that how one functionality of system works with other when they are kept and run in parallel, in order to check the overall functionality of system by observing output results according to some given inputs, black box testing comes for the aid. But this type of testing technique has nothing to do when it comes to finding and removing the specific integral parts of the code, therefore White Box Testing known as WBT is used where code and modules are tested in very much detail.

When considering all of this it is very clear that testing is a very important step of the development process. By using these testing methods, it is made sure that whether the requirements are fulfilled by the system or not. The test cases and the scenarios are analysed by giving input opposing to every module and the component of the system. The test cases of the python logic and code have been executed on Visual Studio Code and the test cases for the web have been executed using Google Chrome, Mozilla Firefox etc. Test cases are very important in discovering the defects, errors and bugs of the system and they point out the components and the modules of the code that are not working properly therefore changes can be made in the code in order to correct mistakes.

6.1 Test Bed

A test bed is an execution environment used for the purpose of testing. A Test bed usually consists of specific collection of hardware, software, a network, an operating system, the product that is being tested, other system software and application software. All the functionalities performed by the system including the modules are tested on this test bed. The test bed used for the testing of “Contextual News Information Retrieval” is a Laptop running Windows 10 running, 8GB RAM and 1 Terabyte Hard Disk, and the applications

used in testing are visual studio code and Google chrome. In user environment by executing and performing the functionalities ourselves through user perspective and analysing the navigation and usability.

6.2 System Test Case

The system test cases are written and executed to find all the possible errors and bugs in the system. System testing supports to improve the functionality performed by the system and it also helps to ensure and validate that the system is being made according to the requirements of the user. Throughout the testing of the system it have been found a large number of errors and bugs and that is exactly what the goals were, many of the bugs were corrected but many were very hard to settle there for they have been contained and minimized.

6.3 Test Cases

Test cases are the scenarios or conditions which are used by a tester to ensure that the system that is being developed is in accordance with the requirement of the user and to detect the defects and the errors in the system before launching it in the market. Test cases help to ensure that the system is error free and performing its functionality without any risks involved, the test cases have aided the testers to detect errors and have aided the development team to remove these errors. Test cases have also highlighted some of those defects which were very hard to deal with and remove. Therefore, to deal with such defects the development team has done some changes to do their containment and to minimize their impact on the system.

Test cases consist of test case ID, reference of use case, QA test engineer, name of person, date of test case, revision history, testing environment, objective, assumptions, steps, status of a test case and preconditions. Test cases for all the features of the system are given below in very much detail.

6.3.1 Sign in

The table 6.1 shows that the sign in of the user is a necessary functionality of the system. Without this feature the user is not able to perform some important functionalities that are related with the system. The most important functionalities include, maintaining of search history, saving of news, news that is based on interest. This test case is triggered when the user enters specific data by itself.

Table 6.1 Test case for sign in

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-1 (sign in)	Tester	Annas Israr
Test Date	15-01-2020		
Revision History	None		
Objective	To sign in or log in the user		
Environment	User Mode		
Assumptions	User is present at sign in page and has registered before.		
Pre-Requisite	System is in functional state.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2.	Display sign in page Input the data to sign in (Email and Password)	Page displayed The user has input the data.	
Comments: sign in without any delay			
Status	☑Pass ☐ Fail ☐Not Executed		

6.3.2 Sign up

The table 6.2 shows that the important feature in the system is to register the user that sign up. The user of the system registers themselves on the system by using this feature. This test case is triggered when the user enters specific data by itself.

Table 6.2 Test case for sign up

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-2(sign up)	Tester	Faizan Ahmad
Test Date	15 -01-2020		
Revision History	None		
Objective	To register new users		
Environment	User Mode		
Assumptions	User is present at signup page		
Pre-Requisite	System is in a functional state.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2.	Display signup page Input the data of the user on system	Page displayed The user has input the data.	
Comments: Signup functionality performed without any delay			
Status	☑Pass ☐ Fail ☐Not Executed		

6.3.3 Search news

The table 6.3 shows that searching the news is one of the main functionalities of the system. This feature allows the user to search the news according to the query of the user and system extracts the news information and bring the result to the user.

Table 6.3 Test case for search news

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-3(Search news)	Tester	Zain Ul Abedin
Test Date	20-05-2020		
Revision History	None		
Objective	To search the desired news		
Environment	User Mode		
Assumptions	User is signed in		
Pre-Requisite	System is in a functional state.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2. 3.	Display home page Input the search query Search event triggered by click	Page displayed The user has input the data. User has clicked the search button	
Comments: news retrieved based on the context with 15-20 seconds delay			
Status	☑Pass ☐ Fail ☐Not Executed		

6.3.4 Search history

The table 6.4 shows that the search history is a very important feature and can be used only by the registered users, the guest users cannot use this feature, this feature helps to maintain and manage the history of the search quires that the user made and provide the user with a list of search queries he had made.

Table 6.4 Test case for search history

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC4 (search history)	Tester	Yaldram Shahzad
Test Date	21-05-2020		
Revision History	None		
Objective	To view the queries of search of the user		
Environment	User Mode		
Assumptions	User is signed in		
Pre-Requisite	System is in a functional state and the user is registered.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2.	Display dashboard page See under the search History	Page displayed History displayed	
Comments: Updated history list of the query is displayed at the dashboard page.			
Status	☑Pass ☐ Fail	☐Not Executed	

6.3.5 Select interests

The table 6.5 shows that select interest is a very important feature and can be used only by the registered users, the guest users cannot use this feature. The signed in user selects its interests according to which it wants its newsfeed to be updated.

Table 6.5 Test case for select news interests

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-5(select interests)	Tester	Faizan Ahmad
Test Date	25-05-2020		
Revision History	None		
Objective	To select the topics of interests from mentioned categories by the system		
Environment	User Mode		
Assumptions	User is signed in		
Pre-Requisite	System is in a functional state and the user is registered.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2.	Display dashboard page Select the interest of news you wish to see	Page displayed Interests selected and saved in the database	
Comments: dashboard is refreshed, and newsfeed is updated			
Status	☑Pass ☐ Fail ☐Not Executed		

6.3.6 Save news

The table 6.6 shows that save news is a very important feature and can be used only by the registered users, the guest users cannot use this feature. The signed in user save the results of the search that he has made locally to their device in a readable form.

Table 6.6 Test case for save news

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-6(Save new)	Tester	Annas Israr
Test Date	25-05-2020		
Revision History	None		
Objective	To save the results of the search to a local drive		
Environment	User Mode		
Assumptions	User is signed in		
Pre-Requisite	System is in a functional state and the user is registered.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2.	Search news Press the save/download/export button.	Search result displayed A CSV file is downloaded to local drive.	
Comments: saved the contents of the search in a readable file			
Status	☑Pass ☐ Fail ☐Not Executed		

6.3.7 Latest news

The table 6.7 shows that latest news is a very important feature and can be used by the registered users and the guest users. This feature allows the system to display the latest news events that are happening throughout the world to the users.

Table 6.7 Test case for latest news

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-7(latest news)	Tester	Yaldram Shahzad
Test Date	30-05-2020		
Revision History	None		
Objective	Display the latest news		
Environment	User Mode		
Assumptions	User is signed in		
Pre-Requisite	System is in a functional state and the user is registered.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2.	Open landing page Click on the latest news button	Page opened Latest news of interests displayed	
Comments: The news is displayed of the latest topic.			
Status	☑Pass ☐ Fail ☐Not Executed		

6.3.8 Newsfeed

The table 6.8 shows that newsfeed is the most attractive feature of the system which cannot be accessed by a guest user, the user must be signed in to use this functionality. This feature displays a list of news to the user which are according to the interests that he had selected

before and according to its past searches. This is a modified information retrieval to the user.

Table 6.8 Test case for newsfeed

Test Case ID	**Use Case Reference**	**QA Test Engineer**	**Name of Personnel**
TC-1	UC-8(newsfeed)	Tester	Yaldram Shahzad
Test Date	05-06-2020		
Revision History	None		
Objective	Display the news of the user's interests		
Environment	User Mode		
Assumptions	User is signed in and has selected the interests		
Pre-Requisite	System is in a functional state and the user is registered.		
Steps #	**Execution Description**	**Procedure Result**	
1. 2. 3.	Open landing page Click on the newsfeed News according to selected interest displayed	Page opened Newsfeed page opened See the news of yours interests	
Comments: the news is displayed of the interests selected before			
Status	☑Pass ☐ Fail ☐Not Executed		

6.4 Summary

In this chapter testing, analysis of test results and validation of the system have been done. First, in testing, it is an activity or process in which errors are discovered and then those errors are removed from the system thus that the system contains no defects and performs its functions correctly. To achieve this part of goal unit testing, integration testing, system testing, black box testing and white box testing have been used. All the specific parts of the system have been evaluated and tested therefore unit testing has been done. Integration testing and white box has been a continuous procedure and have been done in parallel to development and after compilation along with black box testing.

The errors or the testing results that were faced are then analysed and corrected some of the errors were very difficult to deal with and to correct, so that is why they were contained and minimized.

After the correction of these errors are then validated the system to make sure that the functionalities performed by the system are working fine and the system is doing what it is supposed to do and fulfilling the requirements.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.0 Overview

In this chapter, the conclusions and future recommendations of the system, how much functionalities according to the ones discussed earlier in the system, are successfully implemented, what milestones are achieved and what future work can be done for the sake of improvement. Milestones achieved in developing the system are discussed briefly. Limitations of the project are also discussed that needs to be handled in future if the system is upgraded.

7.1 System Overview

"Contextual News Information Retrieval" is a web-based application for searching news as required with ease. Users can register themselves to access all the features of the system. A user can search its desired news over the internet through using the system, users can view their search history, see interest-based news, latest news, and can download the information to their system. The thing that differentiates this system from other existing systems of this type is that this system can search for the required news in any format, although it is an image file or an E-Newspaper.

7.2 Milestones Achieved

Milestones testify the completion of a project from its starting to end date. They identify what functionalities or features have been implemented. The milestones that are achieved successfully are given below:

7.2.1 User Registration

Users can register themselves successfully by providing a few details about him, that include their first name, last name, email, and password. After registration, the user can enter its credentials and get authenticated.

7.2.2 News Search

After successful login, user can enter its desired query in the search box. The system performs the context annotation on its entered words to get the context of what has been written, after this the application first looks for the required data in the information already saved and finds the similarity, if the similarity of the news is more than 50% that is the similarity threshold have been set, then it displays the information. If it fails to find the required data, then it searches over the web to provide the user with the results he is asking for. This functionality has been successfully implemented.

7.2.3 Display of Information

This feature is about how the system shows the searched information. This feature has been successfully implemented and the system shows the searched results to user in an understandable format.

7.2.4 Latest news

This feature is about the latest trends that are happening around. The system has this feature successfully implemented and it shows the latest trending topics to both (logged-in user and guest user).

7.2.5 Newsfeed

This feature was intended to show the user news according to its selected interests. The system now successfully shows the news about topics that the user had selected, from the five websites that were mentioned earlier.

7.2.6 Search History

This feature was intended to save the user's search history. It has been implemented successfully and the system is maintaining the search history of the registered user.

7.2.7 News Interests

This feature was about the user's selection of interests from the following four categories that include: Sports, Entertainment, Business, and Pakistan. The system now allows the user to select its interests from these four categories and provide newsfeed according to these interests.

7.2.8 Save Information on local machine

This system feature was intended to allow the user to download the information on its system. This functionality has been successfully implemented and users can now download the information.

7.3 Limitations

Although the system has implemented the features that were proposed, but still there are some limitations that emerged from discussions. Some limitations of this web application are that the application is currently limited to five news websites, only four categories available for users to select as their interests are sports, entertainment, Pakistan, business and the sources that are covered by the system are Pakistani news websites.

7.4 Future Work

In general, if talking about a news retrieval system then it should be able to retrieve any kind of news from any source, but still this system facilitates many researchers, journalists and student to get their required data with minimum effort. If mainly talking about expanding the scope of the application to more sources and websites, it can be time-consuming as separate scrappers need to be designed for each website and each template.

The possibility of implementing a full-fledged news retrieval system with a broader scope is still an open debate that can be worked on in future. News retrieval systems have a large amount of data as something is happening around in the world with each passing second. So, a more dynamic system can be developed that can handle any website coming its way with ease.

7.4 Summary

In this section, the findings of the project were discussed as now the aim to close the project. The limitations that came in the way of developing the system are also discussed. Milestones are something that trace the completion of a specific work and so all the milestones that were identified in the initial stages of development are successfully achieved. A user can now register/log-in to the system, can view its newsfeed according to the interests he selected. User can download the information successfully. A very important portion of the system was searching the news according to users query, which has also been implemented successfully. System now understands the context of the user's entered text and can successfully find the data related to that in the saved information with 50% similarity threshold with 15-20 seconds delay, or over the web. Search history of the user has been maintained. Limitations that were on the way to developing the system were also discussed in this chapter, these limitations can be addressed in future versions of the project. There is some future work that was discussed and can be done with advancement in available libraries in future. Future recommendations were also discussed.

REFERENCES

1. “Google News, ” Google, 22 September 2002. [Online]. Available: https://news.google.com. [Accessed 2 January 2020].
2. “ABC News, ” Walt Disney Telivision, 15 June 1925. [Online]. Available: https://abcnews.go.com/. [Accessed 2 January 2020].
3. “Concept: Use Case Model,” UTM, [Online]. Available: http://www.utm.mx/~caff/doc/OpenUPWeb/openup/guidances/concepts/use_case_model_CD178AF9.html. [Accessed 1 June 2020].
4. “What is Unified Modeling Language (UML)?, Visual Paradigm,Online]. Available: https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-uml/. [Accessed 1 June 2020].
5. “Tesseract OCR,” Google, [Online]. Available: https://opensource.google/projects/tesseract. [Accessed 2 June 2020].
6. “Pandas,” [Online]. Available: https://pandas.pydata.org/. [Accessed 2 June 2020].
7. “Beautiful Soup Documentation,” [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed 2 June 2020].
8. “Requests,” [Online]. Available: https://requests.readthedocs.io/en/master/. [Accessed 18 06 2020].
9. C. Academy, “Code Academy, ” [Online]. Available: https://www.codecademy.com/articles/scikit-learn. [Accessed 17 june 2020].
10. S. Prabhakaran, “Machine Learning Plus,” [Online]. Available: https: //www.machinelearningplus.comnlp/cosine-similarity/#:~:text=Cosine%20similarity%20is%20a%20metric,in%20a%20multi%2Ddimensional%20space.&text=The%20smaller%20the%20angle%2C%20higher%20the%20cosine%20similarity.. [Accessed 17 june 2020].