News Article Summarization

A PROJECT REPORT

Submitted by

BHAVSAR DARSHAN VIJAYKUMAR [19BECE30028]

JETHAVA PRACHI [19BECE30140]

JAIN SHREYAS SUNILKUMAR [19BECE30183]

In fulfillment for the award of the degree

Of

BACHELOR OF ENGINEERING
In
COMPUTER ENGINEERING

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH,
GANDHINAGAR

Kadi Sarva VishwaVidyalaya, Gandhinagar

2022 - 2023

LDRP Institute of Technology and Research

Computer Engineering Department

CERTIFICATE

This is to certify that the Project Work entitled **"News Article Summarization"** has been carried out by **BHAVSAR DARSHAN VIJAYKUMAR (19BECE30028), JETHAVA PRACHI (19BECE30140) AND JAIN SHREYAS SUNILKUMAR(19BECE30183)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering (7th Semester) of Kadi Sarva Vishwavidyalaya University, Gandhinagar during the academic year 2022-23.

Guide:

Prof. Sandeep Modha	**Prof. Sandeep Modha**
Internal Guide,	HOD – CE,
LDRP ITR.	LDRP ITR.

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude and thankfulness towards all those concerned with our project.

Firstly, we are thankful to LDRP-ITR for undertaking this project. We are sincerely indebted to Prof. Sandeep Modha for giving us the opportunity to work on this project. His continuous guidance and help have proved to be the key to our collective success in overcoming the challenges that we have faced during the project work. His support made the project making experience a pleasantly memorable one. Without his help at all stages in spite of his own workload; the completion of the project would not have been possible.

We would also like to express our gratitude to our friends and of course, CE & IT Department of LDRP-ITR.

Last but not least we are thankful to the almighty God and our parents for giving us such a good atmosphere to work hard and succeed.

Regards,

BHAVSAR DARSHAN VIJAYKUMAR
(19BECE30028)

JETHAVA PRACHI
(19BECE30140)

JAIN SHREYAS SUNILKUMAR
(19BECE30183)

ABSTRACT

NEWS ARTICLE SUMMARIZATION

Information on the World Wide Web and in other electronic form is increasing tremendously. The major challenge is to find relevant information from large amount of data. Summaries are often necessary to enable timely relevancy assessments, information extraction, or information analysis from source material. Text summarization is an effective technique that is used in combination with Information Retrieval and Information filtering systems to save the user time. Therefore there is a need for some form of information compression which can be achieved by various mining tasks like classification, clustering and summarization that help in understanding the information. Large amount of web content is news. News websites are daily overwhelmed with plenty of news articles. This project presents an effective approach for single document news article summarization to help people obtain the most important information in the shortest time. Further, an author can also register to our website and publish their article's short summary by themselves only. It will help us in our dataset.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Project Overview

Millions of web pages and websites exist on the Internet today. Going through a vast amount of content becomes very difficult to extract information on a certain topic. Google will filter the search results and give you the top ten search results, but often you are unable to find the right content that you need. There is a lot of redundant and overlapping data in the articles which leads to a lot of wastage of time. Our project summarizes new articles so that it gives meaningful information and also saves time.

1.2 Problem Statement

With the tremendous increase of digitized information, the mining task has become a crucial tool for aiding and understanding the information. This includes clustering, classification, categorization and summarization. The major challenge is to find relevant information from large amount of data. Summaries are often necessary to enable timely relevancy assessments, information extraction, or information analysis from source material. A News articles are often too long to read, all users want is a summary. Readers want to save time but also be aware of what's happening worldwide. Previously, articles were summarized by humans which requires a lot of time. Time can be reduced by automated summarization.

1.3 Objectives

Objective is to create an appropriate machine learning model that generates a meaningful fixed length summary of the news article.For this, We are using a dataset which contains articles and headline pairs from several leading newspapers of the country. By using this data, we can use natural language processing and deep learning to provide a solution. A web application is then built which is integrated with the model built. On web app, authors can login/sign up and submit details of their news articles and this data will be added to dataset.

2. DOMAIN ANALYSIS

2.1 Customer

- News article readers
- Author

2.3 Dependencies/ External Systems

Following are the tools / technologies, on which our system depends for its completion,

Programming language: Python

Front-End: HTML, CSS, JavaScript and Bootstrap

Back-End: Django

Hardware interface: 8GB RAM, WINDOWS 10

Database: SQLite

Tools: PyCharm, Jupyter Notebook

Frame work: MVT

3. REQUIREMENTS ANALYSIS

3.1 Requirements

3.1.1 Functional Requirement

1. **Registration:**
   - To enter into this site user has to register himself first. Requirements of registration are first name, last name, user name, email-id, password, confirm password etc.
2. **User Login:**
   - The System provides facility to login into the system.
   - Enter username and password
   - User Profile page
3. **Add News article:**
   - The user can add their news article on web app.
4. **Forgot Password**
   - The user can send reset link to the mail id to reset password.
   - Input: Email id
   - Output: Reset link send to Email id.
5. **Logout:**
   - The system provides the facility to logout from the site
   - Input : Select logout option
   - Output : Logout from the system
   - Processing : User will logout

6. **Summarization:**

   - User can do summarization by adding complete news article.

3.1.2 Non-Functional Requirement

1. **Performance Requirements:**
   - The system need to be reliable
   - If unable to process the request then appropriate error message
   - Web pages are loaded within few seconds

2. **Safety Requirements:**
   - The details need to be maintained properly
   - Users must be authenticated
   - The database must be kept backed up

3. **Security Requirements:**
   - After entering the password and user id the user can access his profile
   - The details of user must be safe and secure
   - Sharing of details

3.1.3 Data Requirements:

- Minimum 1GB needed to store our database.
- 512MB RAM is also needed to install our whole system.

3.1.4 External Requirements:

How will our system connect to other software/components?

External requirements are following;

- To get important notification through E-mail, user must have to provide and email address.
- News articles urls given by user must be correct and working.

3.2 List of Actors

Following are the actors;

1. **Admin:**
   Admin is responsible for adding, deleting or managing any user through database. Admin is the only one with right to change database by admin login.
2. **Tester:**
   Tester will test the whole system and functionality of ML algorithm.
3. **Developer:**
   Examine and update the system as required.
4. **Authors:**
   Authors can login and publish details of their news articles.
5. **End User:**
   End user can summarize news articles by providing complete news article.

3.3 Constraints

The constraints are;

- Only registered authors will able to add articles to the dataset and manage them.
- Author will get any instant massage through e-mail address not on mobile numbers.
- Every author will have its own private password of his/her account.
- To Reset password, author must have to provide and email address.
- For news article summarization, user must enter complete news article.
- Output of article summarization can not be 100% accurate.

3.4 List of use cases

Following are the use cases;

- **Registration**:

  To enter into this site user has to register himself first. Requirements of registration are first name, last name, user name, email-id, password, confirm password etc.

- **Login**:

  The System provides facility to login into the system. Enter username and password. User profile page.

- **Add news article:**

  Authors can add their news article to the dataset. .

- **Summarization:**

  A Non registered User or Registered user can summarize news article.

3.5 System use case diagram

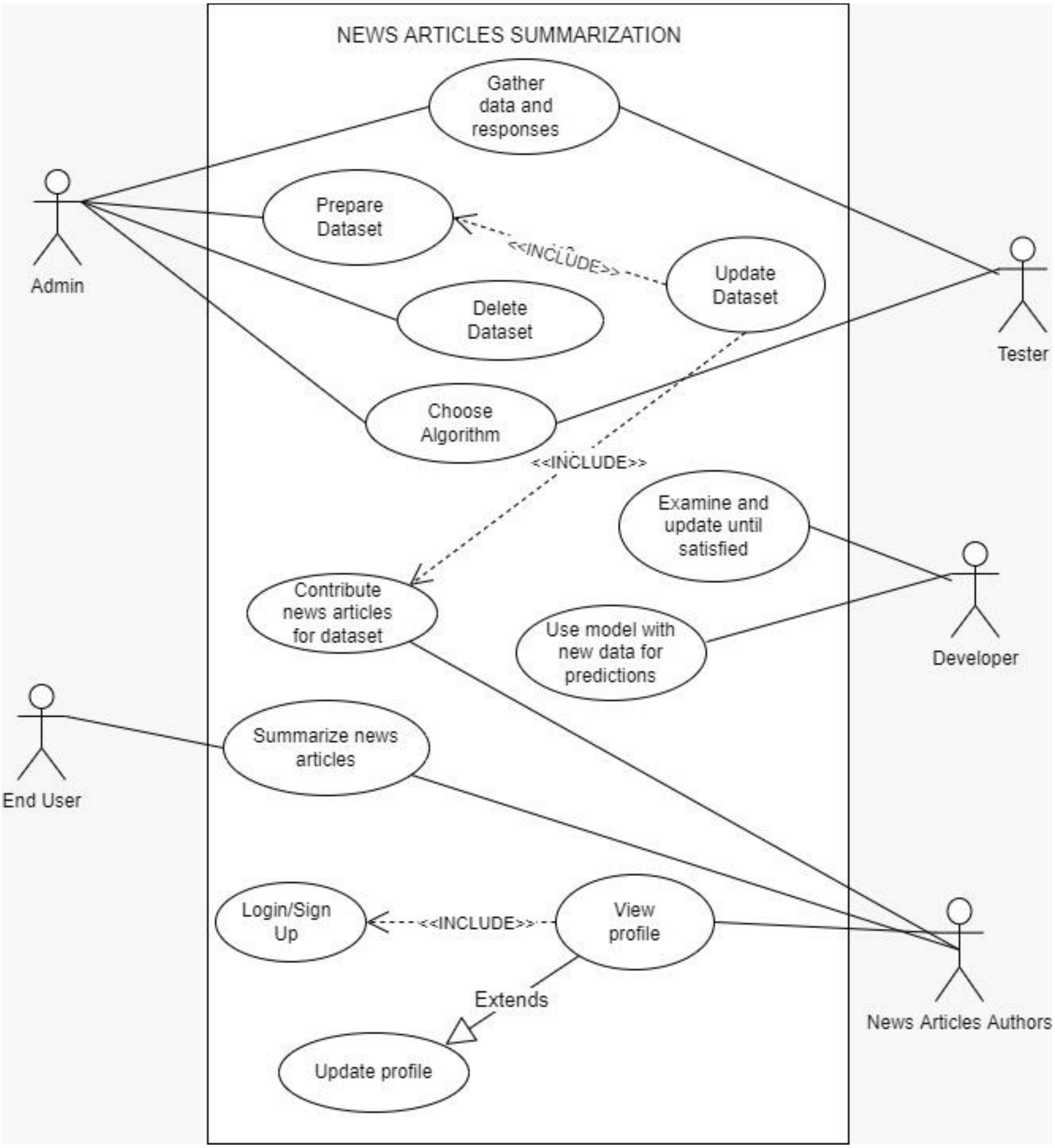
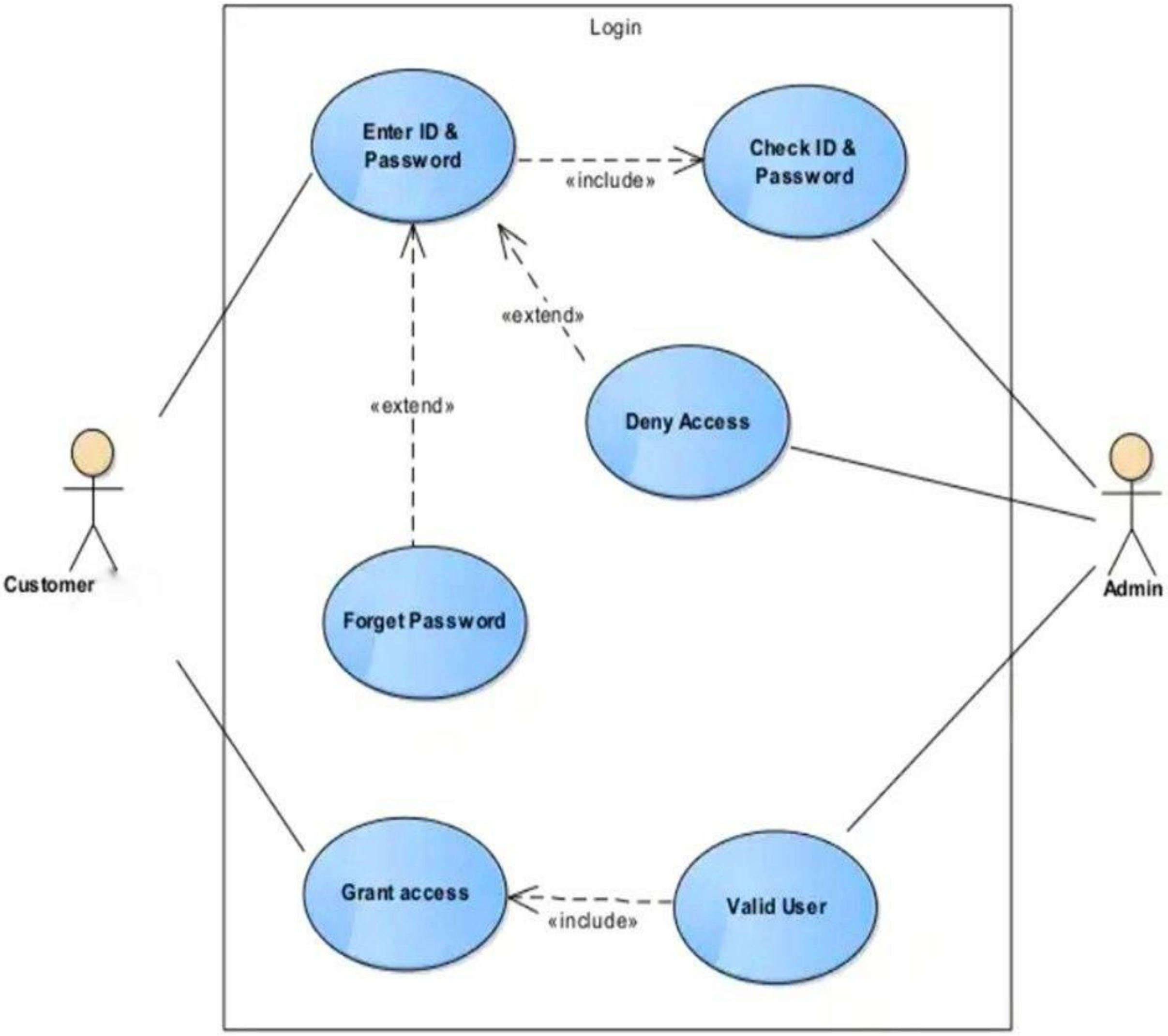


Figure 1 Use case diagram



Use Case Diagram for login page

3.5 Extended use cases

1) Sign up

Section: Main

Name:	Sign up
Actors:	News article's authors.
Purpose:	Sign up to the system
Description:	The user enters his details to sign up to the system.
Cross References:	NONE
Pre-Conditions	NONE
Successful Post Conditions	Sign Up Successful
Failure Post Conditions	Sign Up Failed. Enter correct details.

Alternative Course

Step 1:	The user enters invalid login information
Step 2:	The system displays an error and asks the user to reenter the information.

(2) Login

Section: Main

Name:	Login
Actors:	Administrator, Authors.
Purpose:	Login to the system
Description:	The user enters the username and password to login to the system.

Typical Course of Events			
Actor Action		System Response	
1	This use case begins when a user enters the username and password on the login screen	2	The system validates the information and logs the user into the system

Alternative Course

Step 1: The user enters invalid login information

Step 2: The system displays an error and asks the user re-enter the information

4. DATA FLOW DIAGRAM

4.1 Data Flow Diagram Level 0

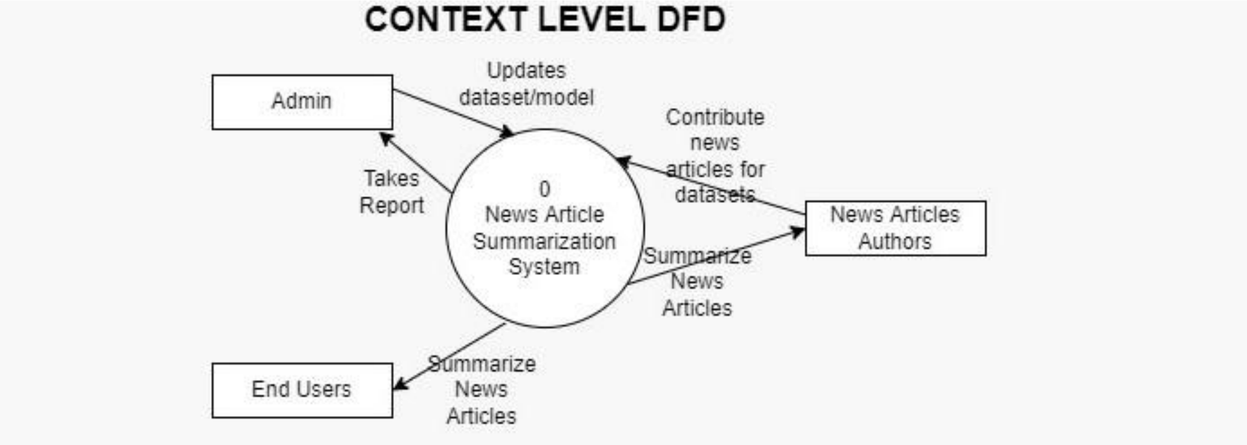


Figure 2 DFD LEVEL 0

4.2 Data Flow Diagram Level 1

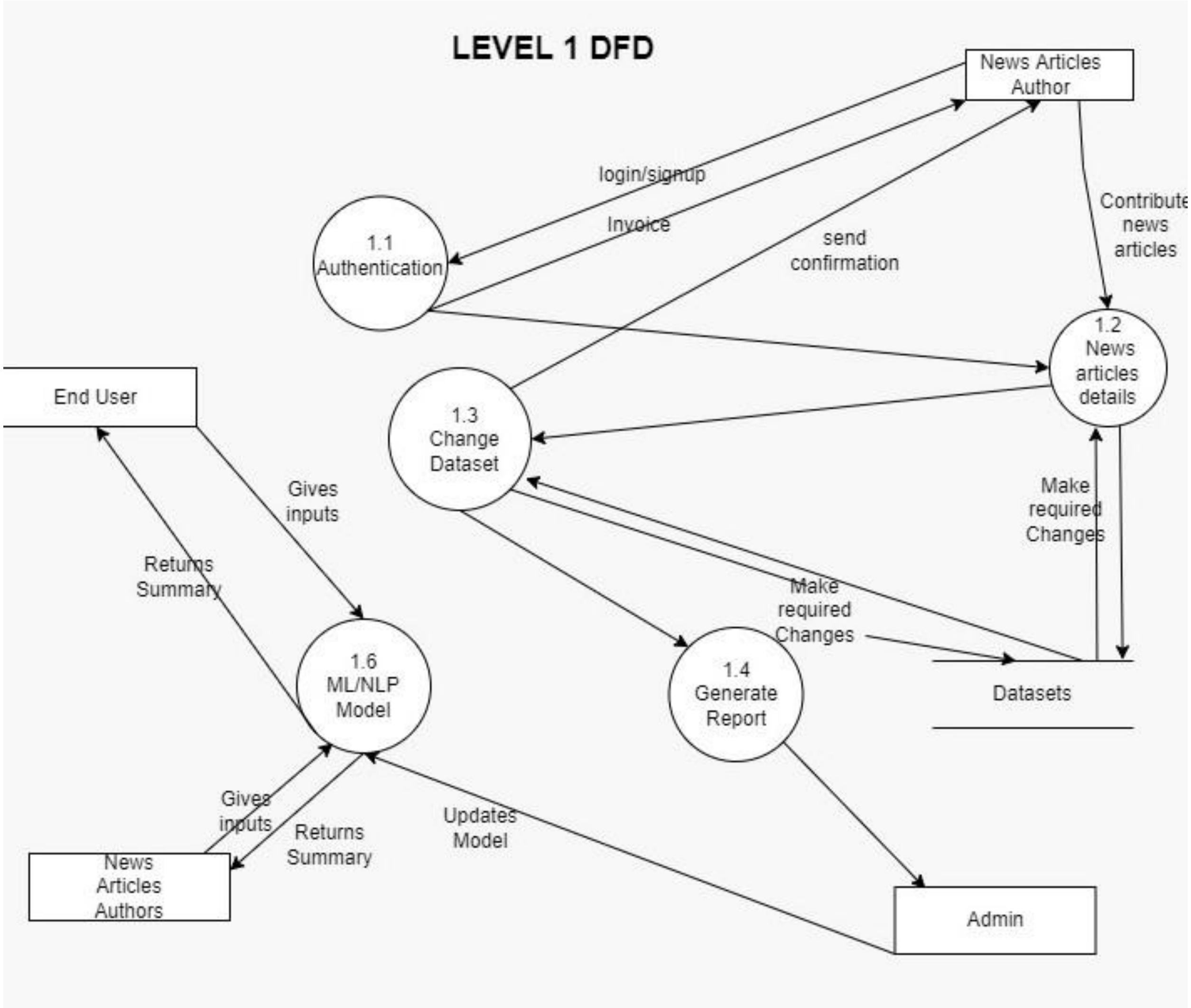


Figure 3 DFD LEVEL 1

4.3 Data Flow Diagram Level 2

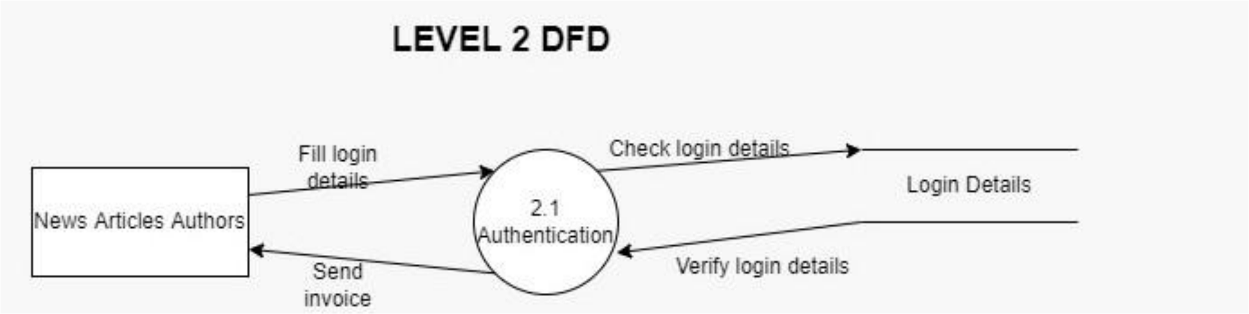


Figure 4 DFD LEVEL 2

5. SYSTEM DESIGN

5.1 System Architecture Diagram

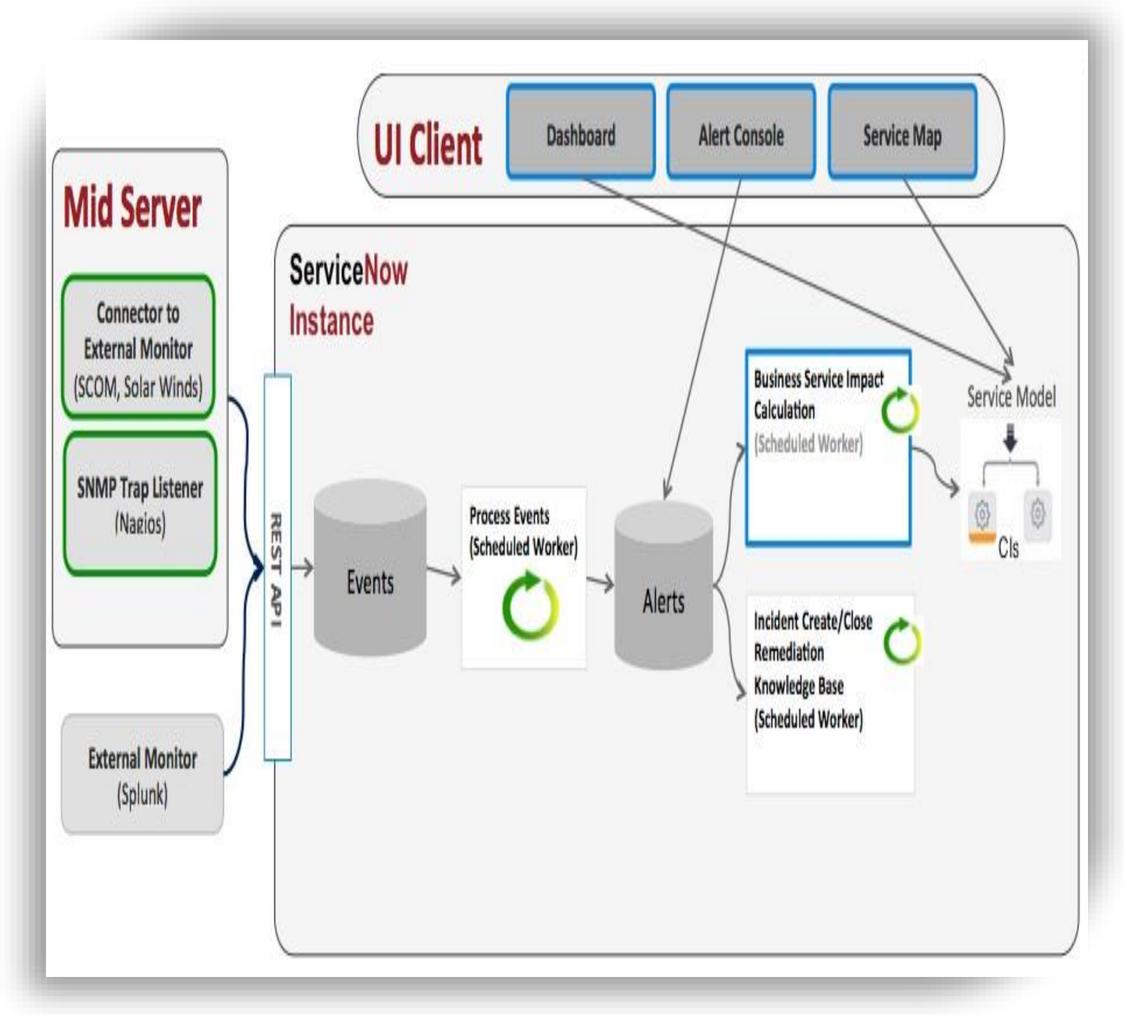

UI Client
Dashboard
Alert Console
Service Map
Mid Server
Connector to External Monitor (SCOM, Solar Winds)
SNMP Trap Listener (Nagios)
External Monitor (Splunk)
ServiceNow Instance
REST API
Events
Process Events (Scheduled Worker)
Alerts
Business Service Impact Calculation (Scheduled Worker)
Incident Create/Close Remediation Knowledge Base (Scheduled Worker)
Service Model
CIs

Figure 5 System Architecture

5.2 Class Diagram

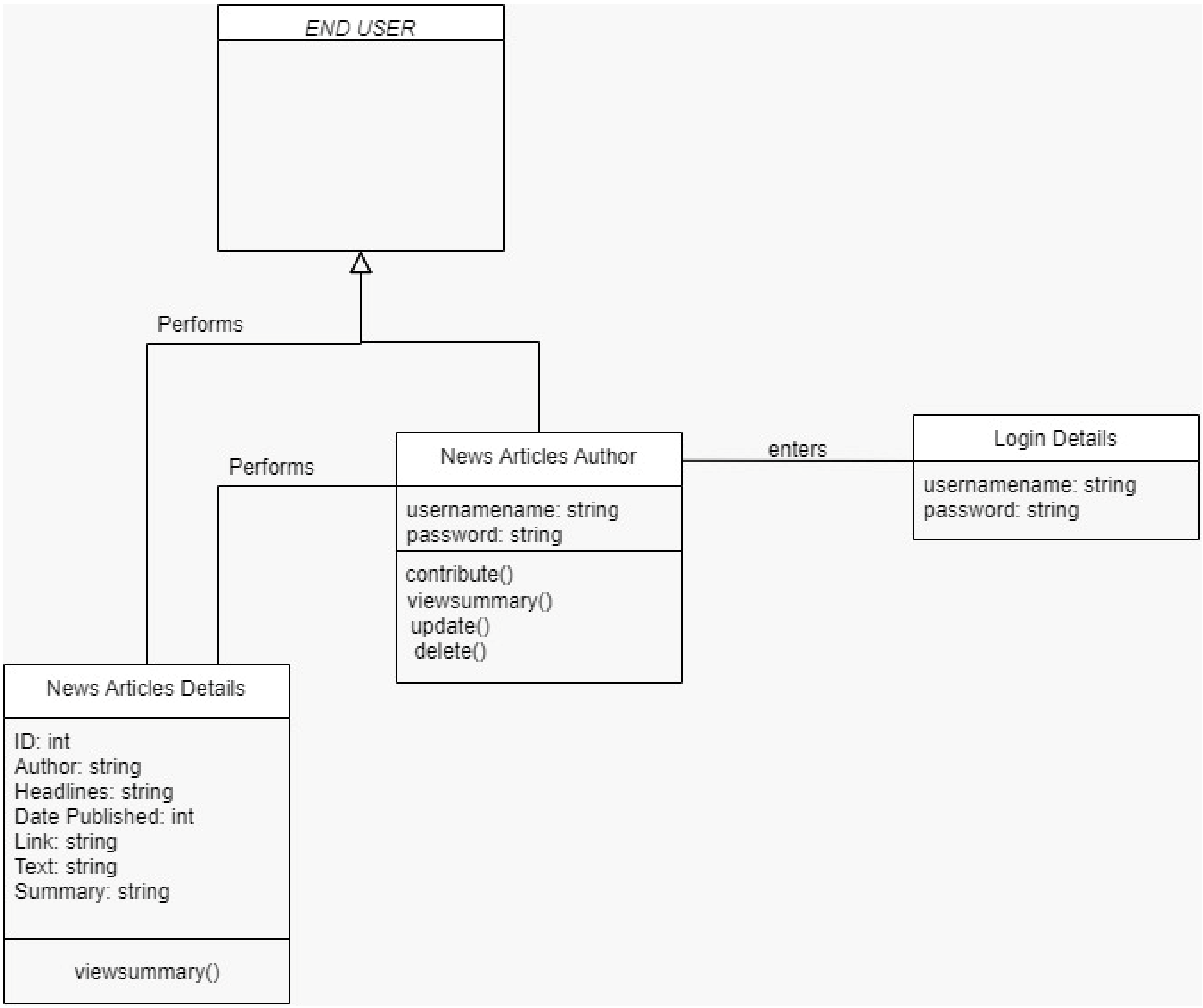


Figure 6 Class Diagram

5.3 Sequence Diagrams

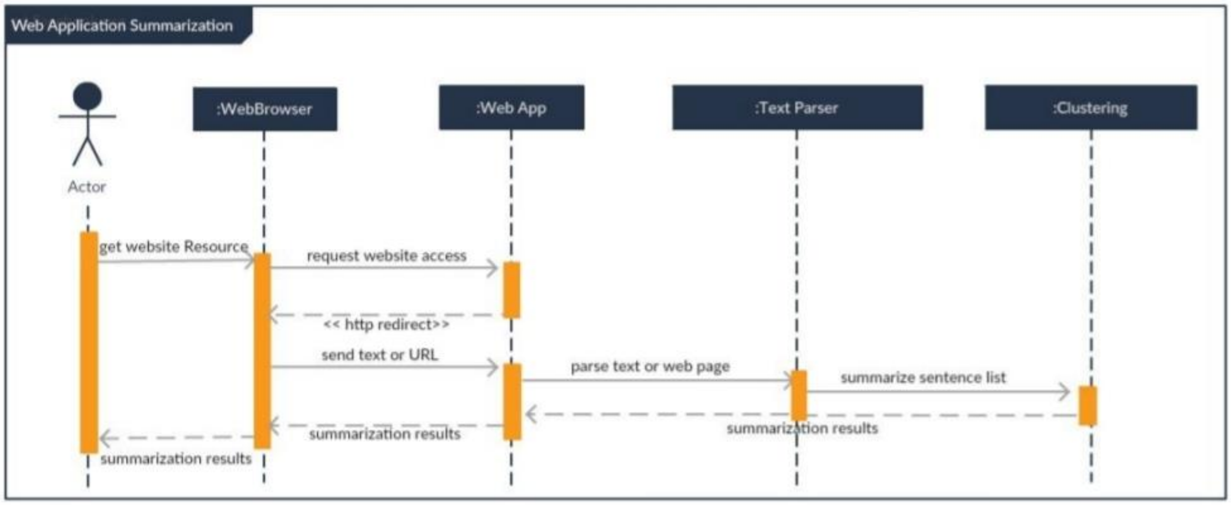
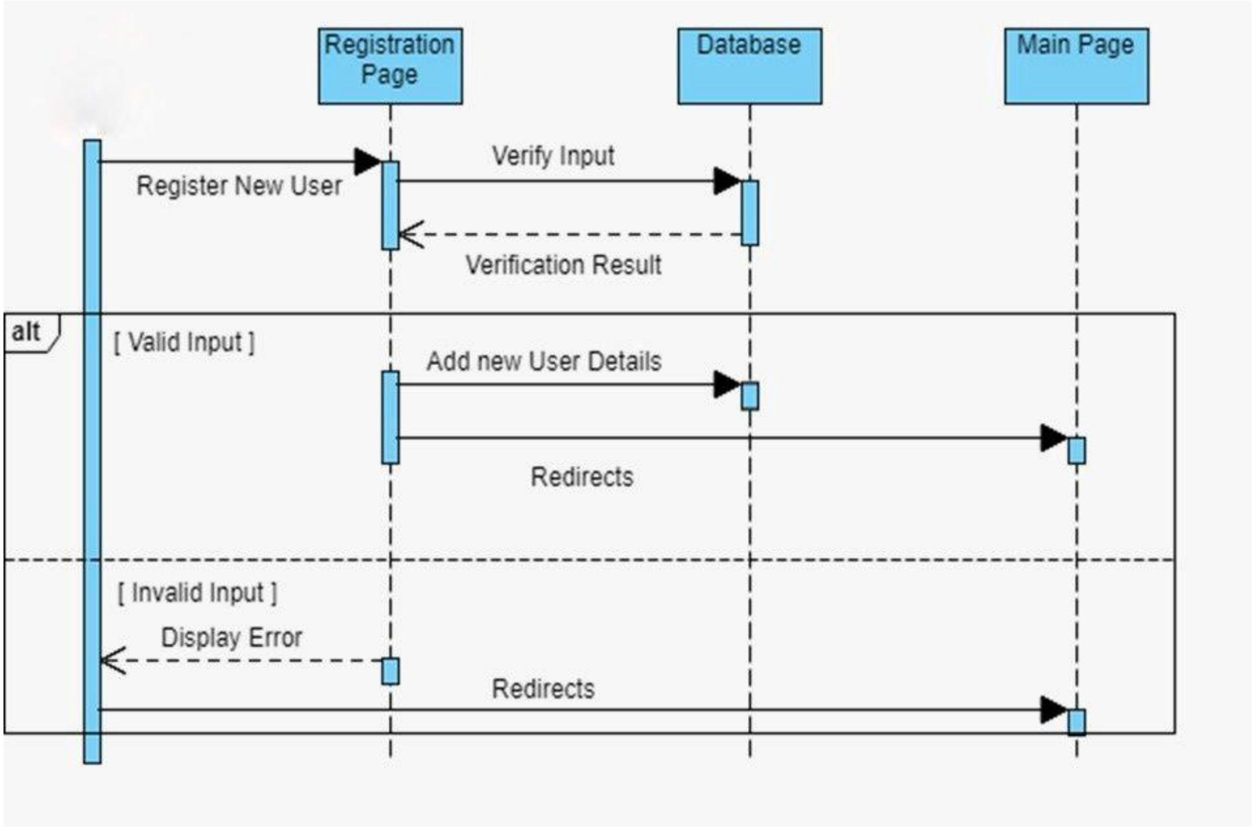
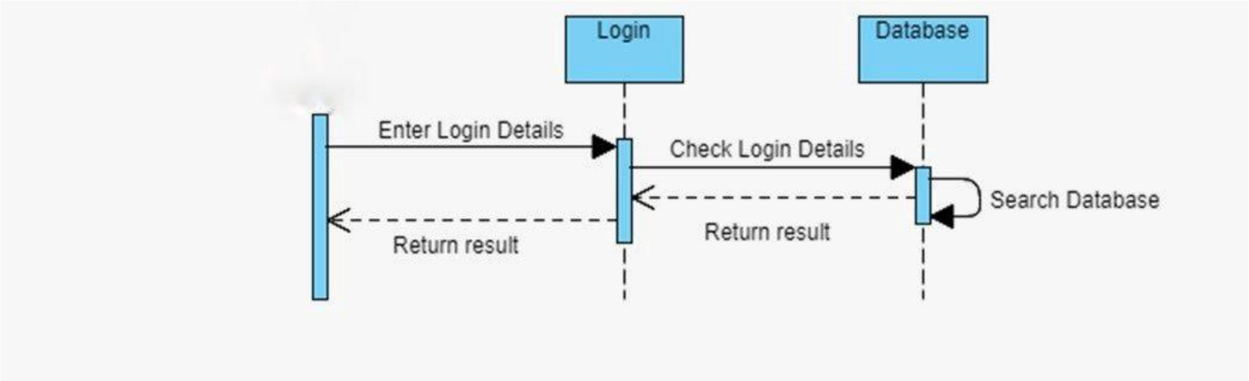


Figure 7 SEQUENCE DIAGRAM



Sequence diagram for Sign up page



Sequence diagram for login page

5.4Activity Diagram

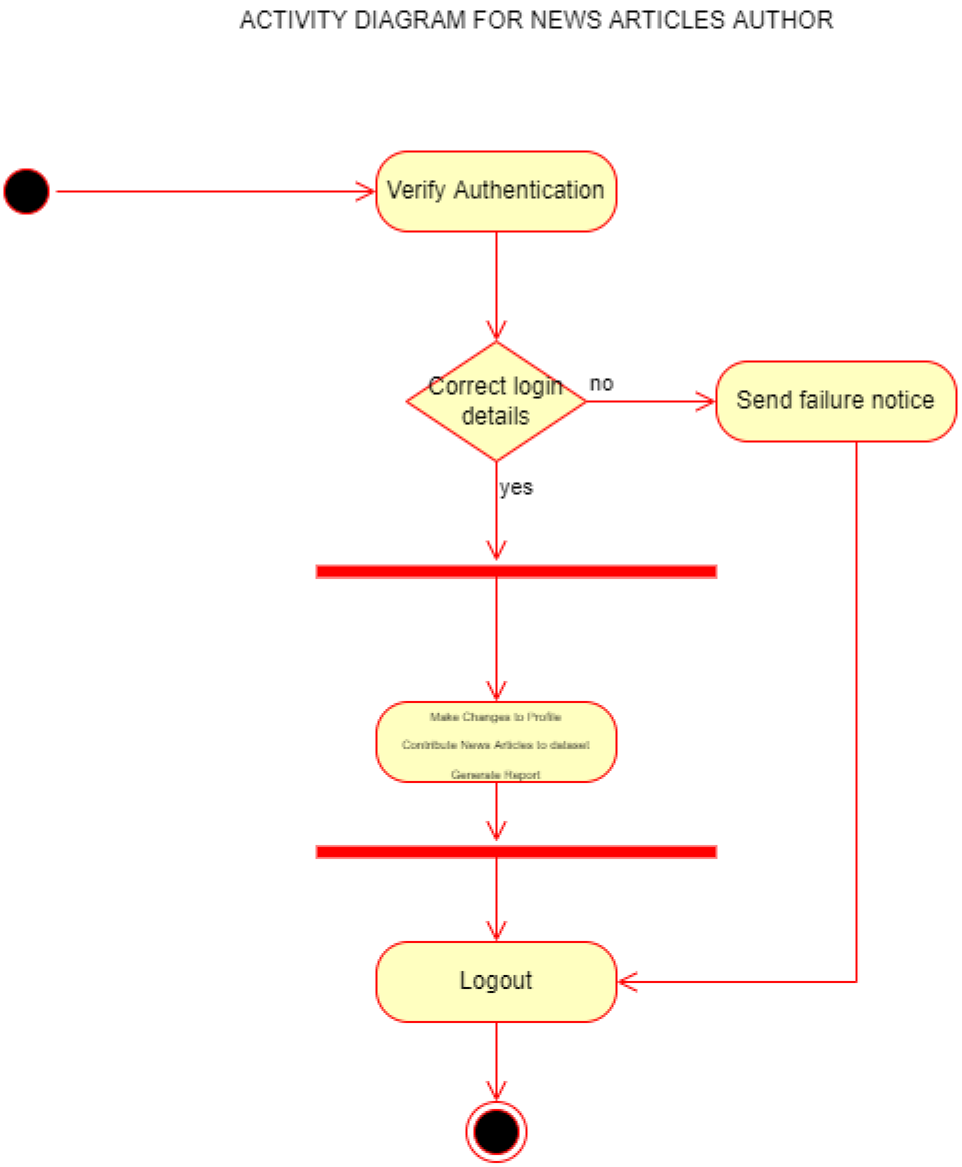


Figure 8 Activity Diagram

5.5 ERD(Entity Relationship Diagram)

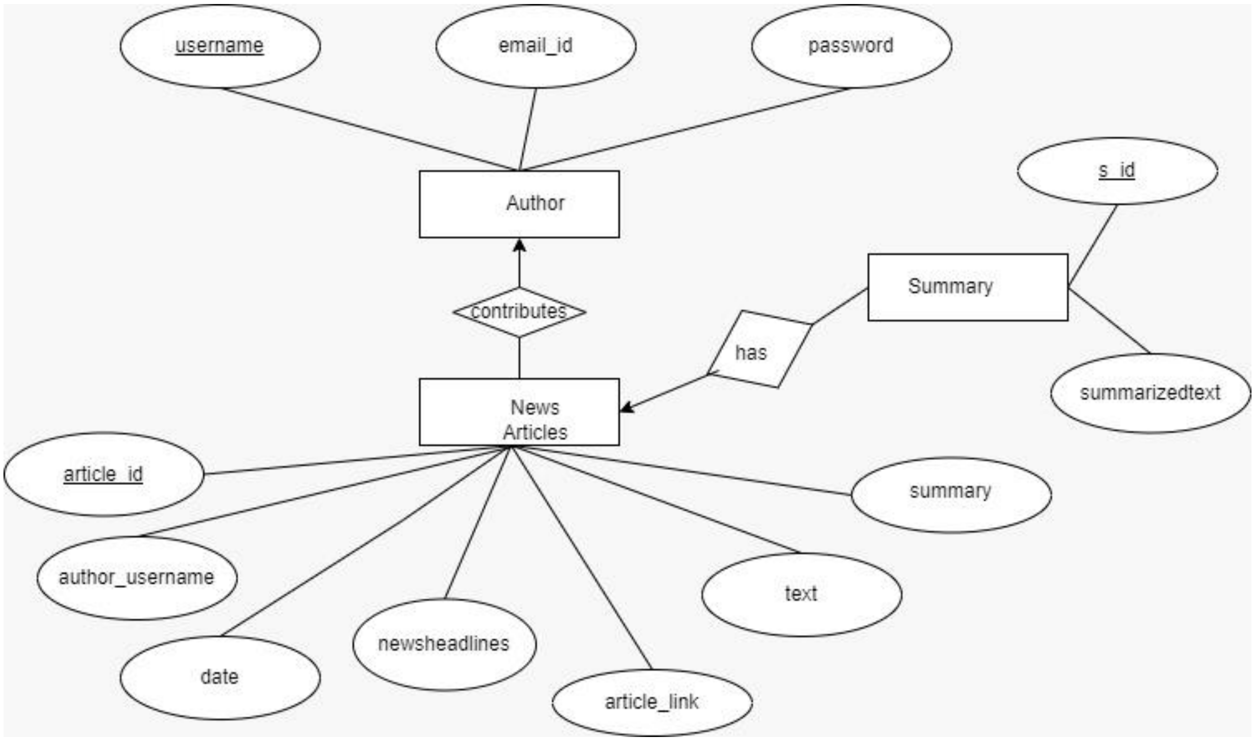


Figure 9 ERD

5.6Data Dictionary

1)User:

Name	Type	Size	Description
ID	Integer	1	Id of the User
First name	String	30	First Name of the User
Last name	String	30	Last name of the User
Username	String	30	Username of the User
Email	String	20	Email of the User

Table 1 Data Dictionary 1

2)Account:

Name	Type	Size	Description
ID	Integer	4	Id of the User
Password	String	30	Password of the User
Email	String	20	Email of the User

Table 2 Data Dictionary 2

3)Admin:

Name	Type	Size	Description
ID	Integer	5	Id of the Amin
Name	String	30	Name of the Admin
Password	String	25	Password of the Admin

Table 3 Data Dictionary 3

6. IMPLEMENTATION DETAILS

This section includes all the implementation details.

6.1 Development Setup

Programming language: Python

Frontend: HTML, CSS, JavaScript, Bootstrap

Back end: Django

Hardware interface: Windows 10

Database: SQLite

Tools: PyCharm, Jupyter notebook

Framework: MVT

SQLite server is an open source relational database management system which is used to store the database using SQL queries. In this project data of halls, wedding lawns and data of accounts are stored in it.

MVT (Model View Template) is used for implementing user interface on computers.

6.2 Deployment setup

As this project is online web based application so we created a github repository and deployed our codebase on github.

6.3 Constraints

6.3.1 Assumptions

1. The client will have appropriate inputs news article summarization.
2. People will have internet connection to approach our web application.
3. Most of the people will visit our website who are authors and want to give their news articles details.
4. Most of the user will be either author or news reader.
5. User may be facilitating for online registration.

6.3.2 System constraints

1. Personal Computer or Laptop
2. Smart Phone
3. Internet
4. Email id

6.3.3 Restrictions

- To Summarize news article, user must have required news article.
- A non registered user will not be able to use publication section.
- News article summarization is using Machine learning so it can not be 100% accurate.

7. TESTING

7.1 Extended Test Cases

Table 5 Test case 1

Sr. no.	Test case	Expected result	Test result
1	Enter Valid Name and Password & Click on Login Button	User should log in.	Successful
2	Enter Invalid Name and Password & Click on Login Button	User should not login.	Successful

Table 6 Test case 2

Sr. no.	Test case	Expected result	Test result
1	Click on Enter news article	Article should be passed to backend	Successful
2	Click on summarize	Article should be summarized	Successful

Table 7 Test case 3

Sr. no.	Test case	Expected result	Test result
1	Click on sign up with valid inputs	New User should be registered in the database	Successful
2	Click on sign up with invalid inputs	User should not be registered to database and error message will be displayed	Successful

Table 8 Test case 4

Sr. no.	Test case	Expected result	Test result
1	Click on add Article	Article should be added and stored in the database	Successful
2	Click on view Article	List of articles should be displayed from the database	Successful

Table 9 Test case 5

Sr. no.	**Test case**	**Expected result**	**Test result**
1	Click on News article summarization	A input form should be displayed	Successful
2	Enter all input fields	Input fields should accept values	Successful
3	Click on Summarize button	Summarized article should be displayed	Successful

7.2 RESULTS/OUTPUT/STATISTICS

7.2.1 %completion.

We have completed our project 100%. We have met most of the functional requirements that we discussed.

7.2.2%accuracy

Article publication system is working 100% accurate. It fulfills all the functional and non functional requirements as we promised. Machine learning model of News article summarization also gives summarized articles.

7.2.3%correctness

As we have tested all the requirements and made their test cases mentioned and clear all the mistakes so now our project is 100% correct.

8. BIBLIOGRAPHY

8.1 CONCLUSION

Our project is only a humble venture to satisfy the needs to manage their project work. Several user friendly coding have also adopted. This package shall prove to be a powerful package in satisfy all requirements of the user. The objective of software planning is to provide a frame work that enable the manager to make reasonable estimate made within a limited time frame at the beginning of the software project and should be update regularly as the project regularly.

At the end it is concluded that we have made effort on following points…

- A description of background and context of the project and its relation to work already done in the area.
- Made statement of the aims and objectives of the project.
- The description of the purpose, scope and applicability.
- We define the project on which we are working in project.
- We describe the requirement specifications of the system and actions that can be done on these things.
- We designed user interface and security issues related to system.
- Finally the system is implemented and tested according to the test cases.

8.2 FUTURE WORK

It can be summarizing that the future scope of the project circles around maintaining information regarding:

- We can add advance features for News article summarization.
- We will host the platform on online servers to make it accessible worldwide
- Integrate multiple load balancers to distribute the loads of the system.
- Implementing the backup mechanism for taking backup on codebase and database on regular basis on different servers.
- We can expand summarization process from one language to multiple regional language.
- Auto-generated questions and MCQs can be created from the given news article.
- Sentiment analysis can be done on given article, which can be further used for tone detection.
- We can scrape news article from different websites in real time and keep their summary on our website.

The above mentioned points are the enhancements which can be done to increase applicability and usage of the project.

8.3 BIBLIOGRAPHY

8.3.1 References

1. https://docs.python.org/3/
2. https://scikit-learn.org/
3. https://www.kaggle.com/
4. http://stackoverflow.com/
5. https://docs.djangoproject.com/en/4.0/
6. https://github.com/