# SHORT NEWS PORTAL

By

**Abdullah Al Shafi**
Roll: 1807004
&
**Ankan Saha**
Roll: 1807060

**Supervisor:**

**Dr. K. M. Azharul Hasan**
Professor
Department of Computer Science and Engineering
Khulna University of Engineering & Technology (KUET), Khulna

KHULNA UNIVERSITY OF ENGINEERING & TECHNOLOGY

Department of Computer Science and Engineering

Report on CSE3200

Course Title: System Development Project

# INDEX

PAGE NO.

## Motivation:

Today, the tremendous information is available on the internet; it is difficult to get the information fast and most efficiently. There are so many text materials available on the internet, in order to extract the most relevant information from it, we need a good mechanism. Text summarization technique deals with the compression of large document into shorter version of text. Text summarizations choose the most significant part of text and create coherent summaries that state the main purpose of the given document. Text Summarization is increasingly being used in the commercial sector such as Telephone communication industry, data mining of text databases, for web-based information retrieval, in word Processing tools. Automatic text summarization is an important step for information management tasks. It solves the problem of selecting the most important portions of the text.

## Objectives:

The objectives of our system are given below:

1) The main objective is to identify the most important information from the given text and present it to the end users.

2) To summarize a text from given raw text.

3) To summarize a text via uploading a text file.

4) To scrap information from website URL or link.

5) To summarize a text providing the link containing the text.

6) To show various categories of news in summarized form.

7) To compare summarizations of the same text using various methods.

8) To do the abstractive summarization of a text.

## Introduction:

A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is ,it reduces the reading time.

Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning. Currently, we enjoy quick access to enormous amounts of information. However, most of this information is redundant, insignificant, and may not convey the intended meaning.

Text summarization methods can be classified into extractive, abstractive and hybrid summarization. An extractive summarization method consists of selecting important sentences from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language. Hybrid uses both of the method to summarize a text or paragraph.

Extractive summarization can be divided into two steps:

1. Pre-Processing step: Pre Processing is structured representation of the original text. It usually includes: a)Sentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence. b)Stop-Word Elimination—Common words with no semantics c)Stemming — The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics.

2. Processing step : In Processing step, features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary.

## Methodology:

At the first, we have started our system development project by implementing necessary algorithms using python in jupyter note book and google colob. After successful implementation, we have started our deployment at PyCharm. In this project, we develop a web application using several python framework like flask(for web development), sqlalchemy(for database management) for backend. Moreover, HTML, CSS, JavaScript, Bootstrap and Tailwind is used to design a responsive web pages.

For Extractive summarization, we have used several algorithm like Term Frequency using SpaCy library as well as NLTK library, Page Ranking Algorithm and sBert model.

For Abstractive summarization, we used Pytorch library to train the Transformers T5 Model using the dataset collected from Kaggle. To train our model in google Colob using GPU accelerator. After saving the model it was loaded in the main project to summary.

For the extraction of data from a website, we took the help of web scraping. This information is collected and then exported into a format that is more useful for the development. To scrap the data, we have used Beautiful Soup (bs4) and request library.

We develop our system in such a way that a user can summarize his/her text in three different ways.

- Providing Raw text.
- Uploading text(.txt) file.
- Providing website link or URL.

Moreover, we have provided various categories of news in our site in a summarized form of original news.

We also implemented a subsystem which can compare text summarization result implemented in different algorithms.

## Algorithm Working Principle:

- ***Term Frequency Algorithm:***

1)*Text Cleaning*: Removing stop words, punctuation marks and making the words in lower case.

- Import spacy library,stop words and punctuation.
- List stopword and add additional next line tag with punctuation.

2) *Word Tokenization*: Tokenize each word from sentences.

- Tokenize the words from the sentences in headline and text using spacy library

3)*Word Frequency Table:* Count the frequency of each word and then divide the maximum frequency with each frequency to get the normalized word frequency count.

- Calculate word frequencies from the headline after removing stopwords and punctuations.
- Calculate word frequencies from the text after removing stopwords and punctuations where tokens in headline have 10 times more frequency.
- Calculate the maximum frequency and divide it by all frequencies to get normalized word frequencies.

4)Weighting Sentences: As per frequency of words in sentences.

- Each sentence is weighed based on the frequency of the token present in each sentence.

- The result is stored as a key-value pair where keys are the sentences in the string doc and the values are the weight of each sentence.

5) *Summarization:* calculate 30% of text with maximum score.

- Take top 30% sentence according to the sequence of the main text.

Psudocode for Term frequency :

```
LET, word_frequencies and sentence_score be two map whose keys are word tokens
and sentence tokens and values are word frequency and sentence score respectively.

1) FOR word in text
        IF word not in stopwords and punctuation
                word_frequencies[word] :=word_frequrncies[word]+1
2) max_frequency := max(word_frequencies.values())
3) FOR word in word_frequencies.keys()
        word_frequencies[word]: = word_frequencies[word]/max_frequency
4) FOR sent in text
        FOR word in sent
                sentence_score[sent]=sentence_score[sent]+word_frequencies[word]
5)  summary := 30% of the text with maximum score
```

❖ ***Ranking Algorithm:***

1) The first step would be to concatenate all the text contained in the articles.

2) Then split the text into individual sentences.

3) In the next step, we will find vector representation (word embeddings) for each and every sentence.

4) Similarities between sentence vectors are then calculated and stored in a matrix.

5) The similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation.

6) Finally, 3 top-ranked sentences form the final summary.

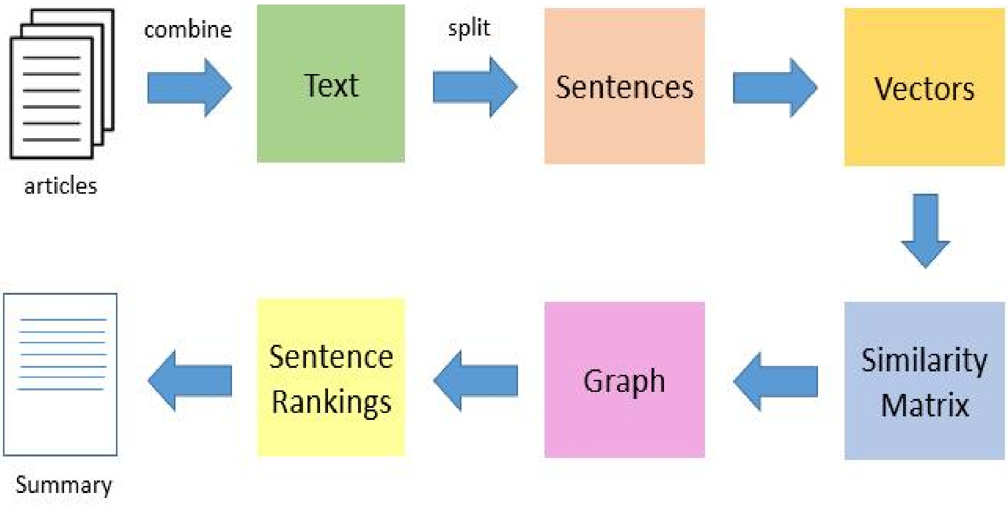

Fig 1: Steps of Ranking Algorithm

## Pseudocode for Ranking Algorithm:

LET Similarity be matrix which store similarity between sentences.

1) Similarity($S_i$ , $S_j$)= $\frac{/\{w_k \mid w_k \text{€} S_i \text{ and } w_k \text{€} S_j\}/}{\log(|S_i|) + \log(/S_j/)}$

2) Apply page rank on similarity matrix.

3) Sort the sentences based on page rank score.

4) Extract top 3 sentences

## UML Diagram:

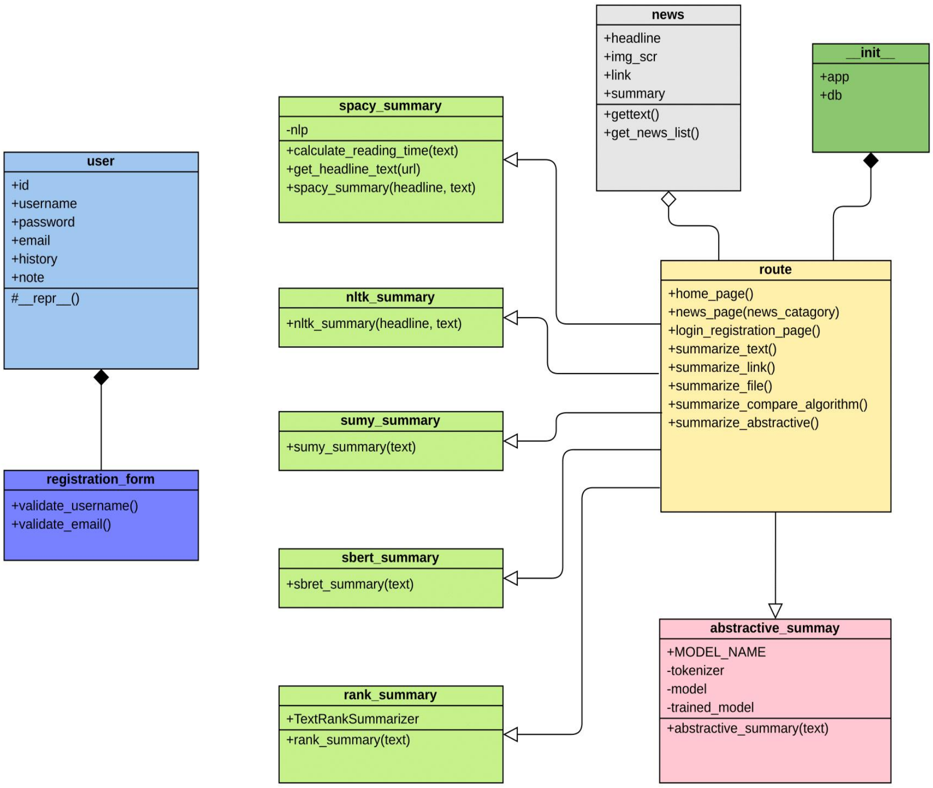

Figure 2: UML Diagram of whole System

## User-Manual:

Our website layout is designed in such a way that any user can easily navigate different section.

- Home Page:
  - Login / Registration Option
  - Select news category
  - Input text giving option for summarization
  - Abstractive Summarization option
  - Contact information of developer
- Login / Registration Page:
  - Login through proper validation
  - Register to our site giving necessary information
- News Page:
  - Get the summarized news
  - Link of the original news
- Summary Page:
  - Provide Raw text and headline
  - Provide URL
  - Provide text (.txt) file
  - Get summarized text
  - Input and Summarized text reading time
- Compare Page:
  - Provide the input through raw text
  - Compare summary of different method like SpaCy, NLTK, Sumy, SBert, Rank.
  - Execution and reading time comparison
- Abstractive Page:
  - Get the Abstractive summary of given Text

## Activity Diagram:

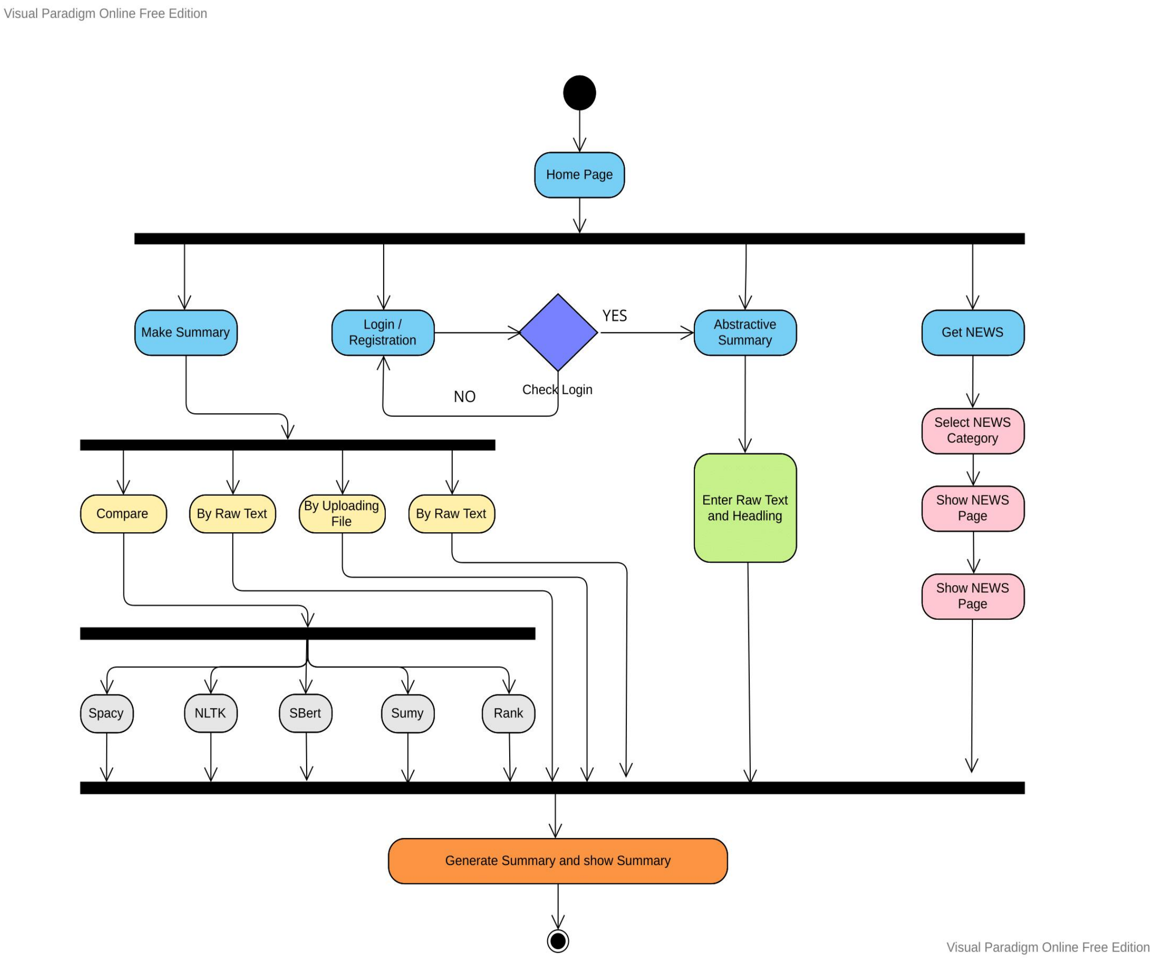

Figure 3: Activity Diagram of User Navigation

## Use Case Diagram:

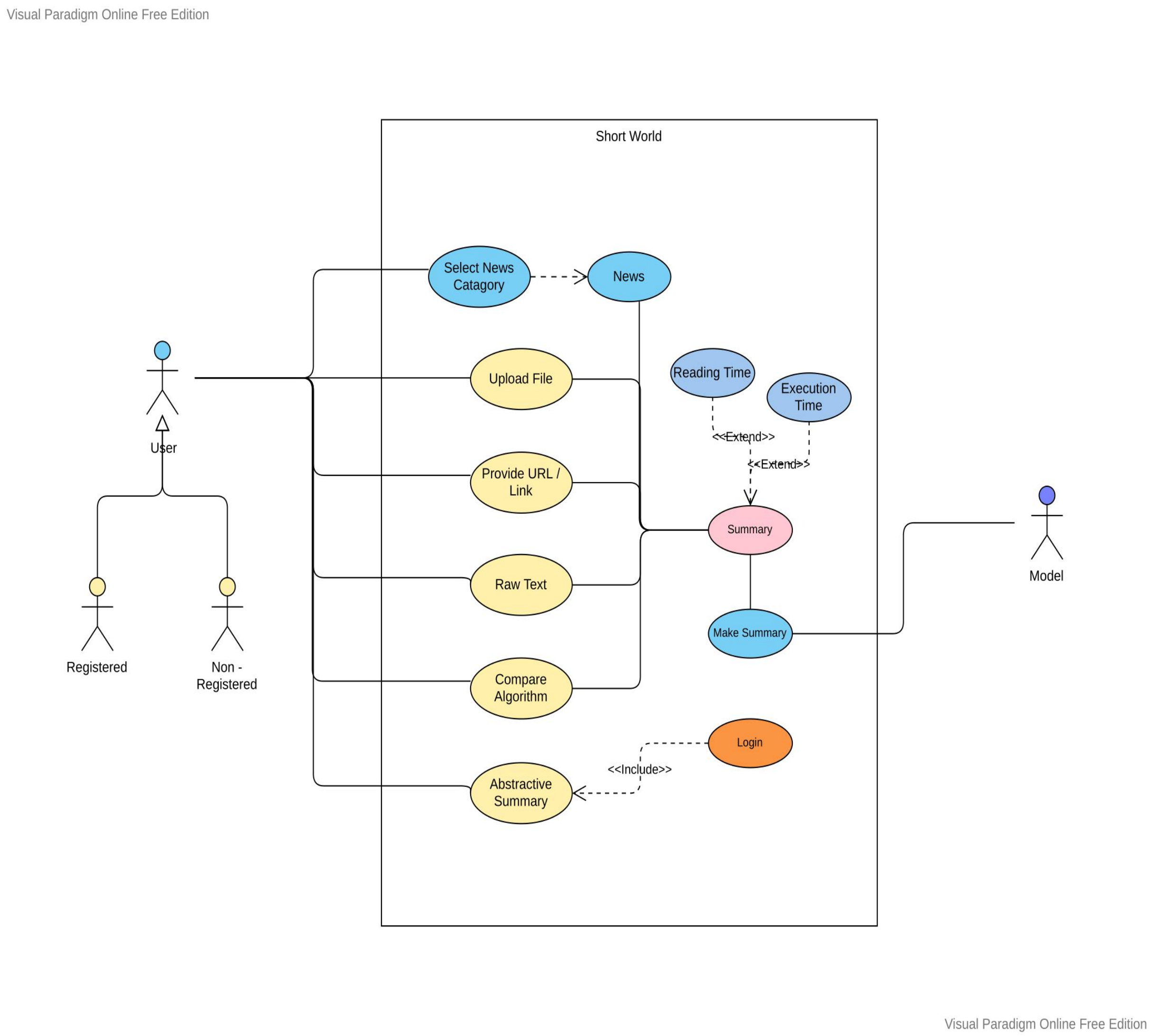

Figure 4: Use Case diagram

## Story Board:

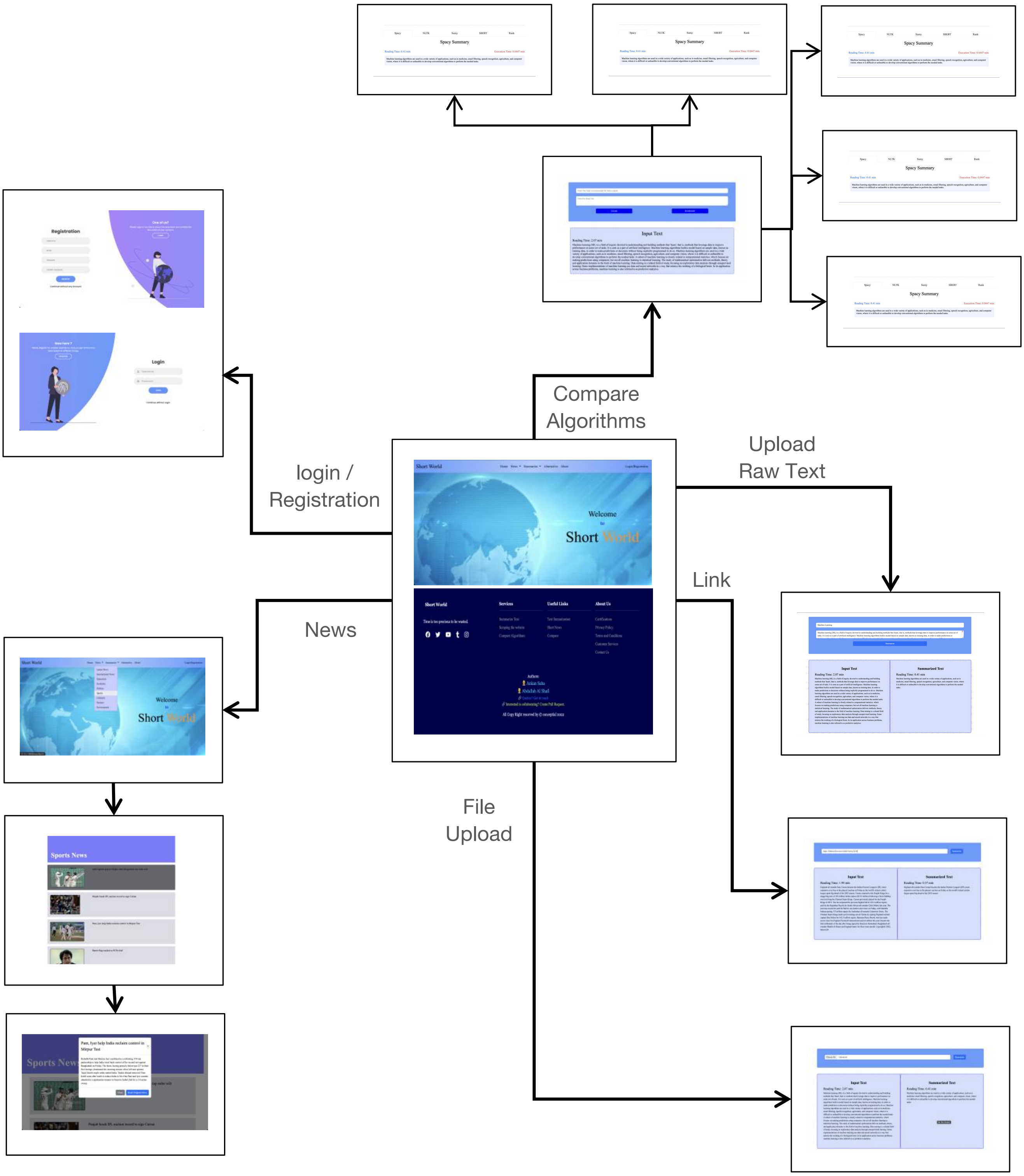

Figure 5: Story board of Website

## Limitation:

As the algorithm outputs the 30% sentences with maximum score, it also shows the references in the output which should not be shown. Again, there may be sentences which sentence score is low because the words in the sentence appear less frequently in the text, so the sentence is not shown in the summarized output. But this sentence may be important to understand the given context properly. Moreover, if the 30% of the sentence tokens is less than 1, then no output will be shown though one or two important sentence can be shown. As many newspapers do not allow scraping some categories of news, those news can't be scraped.

## Conclusion:

The most important advantage of using a summary is, it reduces the reading time. Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning. With the present explosion of data circulating the digital space, which is mostly non-structured textual data, there is a need to develop automatic text summarization tools that allow people to get insights from them easily. Text summarization techniques can be used to read news article shortly. Various tokenization and summarization techniques summarize the same text differently.

## References:

1) https://www.analyticsvidhya.com/blog/2022/02/a-flask-web-app-for-automatic-text-summarization-using-sbert/
2) https://www.thapatechnical.com/2021/06/create-animated-website-using-html5.html#
3) https://dashboard.render.com/
4) https://arxiv.org/abs/1602.03606
5) https://www.youtube.com/watch?v=qtLk2x59Va8
6) https://spacy.io/universe/project/spacy-pytextrank
7) https://online.visual-paradigm.com/
8) https://iq.opengenus.org/textrank-for-text-summarization/
9) https://medium.com/analytics-vidhya/text-summarization-using-spacy-ca4867c6b744
10) https://undraw.co/illustrations
11) https://www.kaggle.com/datasets/sunnysai12345/news-summary?resource=download&select=news_summary.csv
12) https://www.youtube.com/watch?v=KMyZUIraHio
13) https://www.kaggle.com/datasets/sunnysai12345/news-summary?resource=download&select=news_summary.csv
14) https://www.youtube.com/watch?v=9PoKellNrBc
15) https://www.udemy.com/course/complete-data-visualization-in-python/?referralCode=C5022514A150E173DF32
16) https://www.youtube.com/watch?v=uZnp21fu8TQ