Software Requirements Specification Document for project News Aggregator

Alaa Mohamed, Marwan Ibrahim, Mayar Yasser,Mohamed Ayman
Supervised by Dr. Walaa Hassan, Dr Eslam Amer and Eng. Mennatullah Gamil

November 2019

1 Introduction

1.1 Purpose of this document

The main purpose of this document is to outline the description of the news aggregator system. The system aims to crawl relevant news from several ranked sites and platforms based on the user's input news. Furthermore, the document will present the details of the system's functions and features. This document also shows an explicit user interface and how the system will deal with the user interactions. This Software Requirements Specification Document (SRS) is needed for the graduation project at Misr International University (MIU).

1.2 Scope of this project

- The system will:
  1. Aggregate news, articles and tweets all in one place.
  2. Crawl the relevant articles when user input some article to read about it.
  3. Summarize the aggregated information from various sources in one page in order to reduce reading time.
  4. Add Ranking System for the websites agencies/sources in order to avoid contradiction of news
  5. Add Another Ranking system for viewing the most relevant article
  6. Add the Most relevant Media(Images-Videos) through the article according to the website agencies

1.3 Overview

Fast-Text was used at first which is a library for text classification and word embedding which was created by Facebook. So,we are going to use Fast-Text in the system for classification a supervised data set and classifying its text.

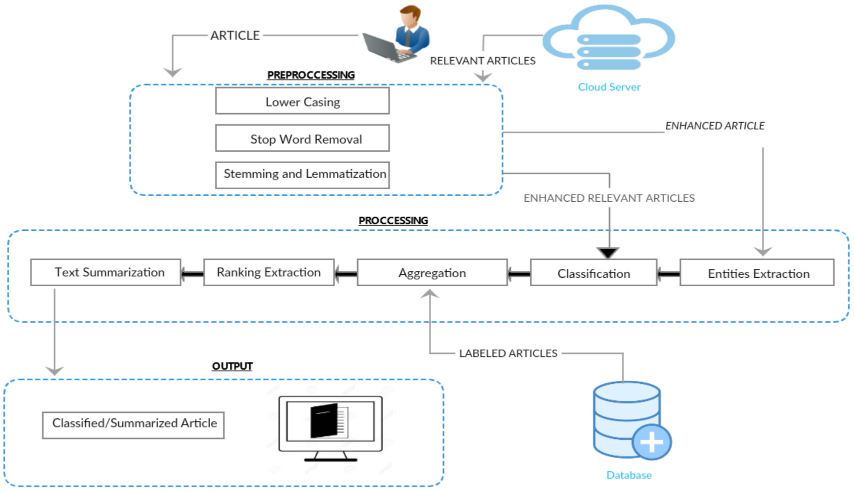


Figure 1: System overview

- Article: When the user enters a new one, it will face many stages in order such as:

  1. The First stage is the PreProcessing Module which consists of:
     - Lowercase: used to reduce the size of the vocabulary in our data that cause multiple copies of the same word meaning.
     - Stop-word Removal: Done to remove small information of a text in order to focus on important words.
     - Lemmatization and Stemming: Remove inflection and map the word into the original/root form.
  2. The second stage is the Processing Module which consists of:
     - Entities Extraction: classifying key elements from a text into predefined categories.
     - Classification: Finding out which label is similar to the input article.
     - Aggregation: Aggregate the related articles from database and sends back to prepossessing for summarization.
     - Ranking Extraction: After having an aggregated content of the input article, we will need to have the ranking of the agencies where we got this content from, in order to take the higher agency if there was any contradiction between the articles.
     - Summarization: Getting the main and important information from all aggregated aticles.
  3. The final state is Output which is:

  - Summarized and well-classified news articles for the user to read.

- Cloud Service: It contains a collection of ranked news articles.
- Database: The labeled articles are saved in it.

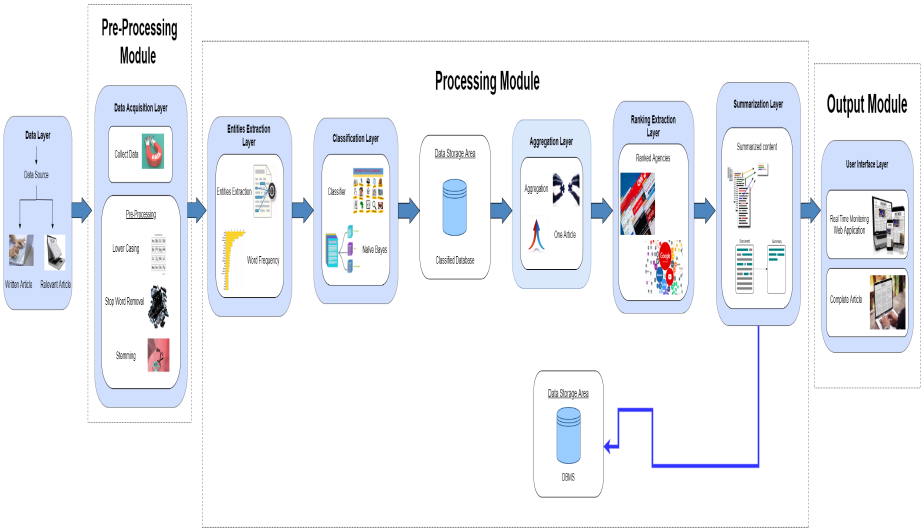


Figure 2: Block Diagram

1.4 Business Context

Online social networks a useful tool for collecting, aggregating and consuming the specific or general contents for various purposes in a certain period of time. An Outsell report (2009), 57 percent of news media clients go to computerized sources,and they are too more likely to turn to an aggregator 31 percent than to a newspaper site 8 percent or other news sites 18 percent.

- Vision:
  - Safe readers' time instead of reading from more than one source.
  - The summarized page contains the required article with the relevant articles all on one page and social media platforms will be placed only in a single location.
- Mission:
  - Develop a news aggregator with machine learning approachable to aggregate relevant articles of a certain input article and summarize all this information on one page.

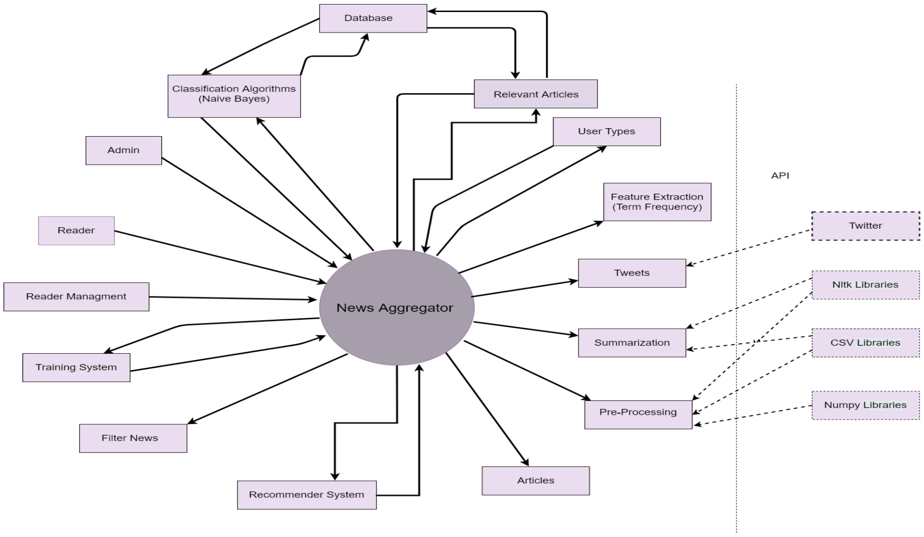


Figure 3: Context Diagram

2 General Description

2.1 Product Functions

The system aims to enhance news aggregation through getting all perspectives in order to give the user all of the possible information.Thus The main goal of this project is to develop a news aggregator with machine learning approach able to aggregate relevant articles of a certain input article and summarize all this information in one page.

2.2 Similar System Information

2.2.1 Atlas: News aggregation service [1]

This system aim to improve the news aggregator platform by merging the RSS news with social media like twitter and news web-services that provides API, it collects data from more than one website and provides people to read articles in any language. In addition of using Twitter API and Reddit API that will help concatenate the news information with the tweets that are written on twitter.

2.2.2 THE RISE OF THE NEWS AGGREGATOR:Legal Implications and Best Practices.[2]

The system aims to make one single platform that collects the pieces of information that are written on different websites.to be easy instead of opening each website and read the news from it. Asthe researchers reach that the internet has become a very important root to know the news.

2.2.3 The Improvement of Indonesian News Curator Classification in Twitter [3]

A similar system that use Twitter user as a news creator which is a valuable client for making news, Two types of news agent were used in this system, the first is the twitter user that tweet the news and the second one is a programmed bot client that aggregate the news and tweets automatically

2.2.4 NewsOne—An Aggregation System for News Using Web Scraping Method [4]

A system that extract the news from multiple sites, newspapers, magazines, and television and merge them in one platform. It is also categorized into many categorize and classify them according to the pieces of information.

2.3 User Characteristics

Expected users are the ones who care about reading lots of news from many sources.They must have a basic knowledge in using computers and must be able to interact with user interface to search for specific articles.

2.4 User Problem Statement

It is hard to the user to get all of the relevant or related articles to an particular article manually, they will consume a lot of time and effort to find the trustworthy or credible source talking about the same topic of the article, hence the summarizing requirement which is a major one in our system surely will reduce user time consumption through exploring events by summarizing all of the aggregated content in just a single page.

2.5 User Objectives

1. Automated system to accurately aggregate news from heterogeneous sources.
2. User can input an particular article to get the other relevant ones.
3. Recording the history of the user in order to know his interests which will be shown to him as news after that.

3 Functional Requirements

3.1 Text Summarization

Table 1: create-frequency-table

Description	This function is done to create a frequency table for text words.
Action	It creates a frequency table for the words in the text after getting rid of the stop words.
Input	Input text article
Output	Table having every word with its frequency in the text.
Precondition	Having an input text without stop words.
Post-condition	The table will have every word with its frequency.
Dependencies	The existence of the text or the article.
Priority	10/10

Table 2: score-sentences

Description	This function is done to give a score to every sentence according to its words.
Action	It gives a score to every sentence by an algorithm which is the division of frequency of every non-stop word in the sentence by the total number of words in the sentence.
Input	Sentences of the text and the frequency table.
Output	The score or the value of every sentence.
Precondition	Having the frequency table and dividing the whole text into sentences.
Post-condition	The score of every sentence is created.
Dependencies	The existence of the frequency table.
Priority	10/10

Table 3: find-average-score

Description	This function is done to find the average score of the sentences.
Action	It gives a score to every sentence by a basic algorithm which is the division of frequency of every non-stop word in the sentence by the total number of words in the sentence.
Input	Sentences Values.
Output	The average score of the sentences.
Precondition	The sentences are done.
Post-condition	Having the threshold or average score of the sentences.
Dependencies	The existence of the sentences with its values.
Priority	10/10

Table 4: generate-summary

Description	This function is done to generate the summary of the input article.
Action	It compares the sentences value with the calculated threshold, if the sentence value is larger, then it will be considered as a part of the summary
Input	Sentences Values, Threshold.
Output	The summarized text.
Precondition	The average score is calculated.
Post-condition	The summarized text is created.
Dependencies	The existence of the sentences with its values and their average score.
Priority	10/10

Table 5:TF-IDF	
Description	This method summarizes article by using NLTK library with TF-IDF formula, by getting the term frequency= (Number of times term t appears in a document) / (Total number of terms in the document) And Inverse documents frequency = loge (Total number of documents / Number of documents with term t in it)
Action	Create frequency matrix then TF-IDF matrix after calculating them and calculating the score with threshold in order to generate the summary
Input	Full News Article
Output	Brief Summarized Article
Precondition	Article in its original form
Post-condition	A Summarized article after calculating the TF-IDF and its threshold
Dependencies	Getting summarized article according to the threshold
Priority	10/10

3.2 Login

Table 6: login	
Description	This function is done to enable the user to get into the system.
Action	It checks if the user and password exists in the database or not.
Input	Email,Password
Output	Boolean true or false depending on the existing of the email and password in the database.
Precondition	The account must be signed up.
Post-condition	None.
Dependencies	The user will not be able to access system features if he is not logged in.
Priority	10/10

3.3 Sign Up

Table 7: signUp

Description	This function is done to create account in the system for the user.
Action	Insert the new user information into the database.
Input	First Name,Last Name,Email, Password,Gender.
Output	Boolean true or false.
Precondition	Check if the user already exists.
Post-condition	The new account is created.
Dependencies	Log in.
Priority	10/10

3.4 Log out

Table 8: signOut

Description	This function is done to get the user of the system.
Action	Sign out from the user's account and return to home page.
Input	None.
Output	Boolean true or false.
Precondition	User must be logged in.
Post-condition	Redirection to home screen.
Dependencies	The user won't be able to use the system until he logs in again.
Priority	8/10

3.5 Encryption-Decryption

Table 9: hashData

Description	Encrypt the upcoming data such as user’s id, password using PHP functions such as (Md5, Sha1).
Action	The System takes the user’s information, and starts hashing all the sensitive data such as Passwords, Ids by converting them into a collection of numbers and characters, for example if the password contains number ’1’, it could be converted to 3k3s in the database.
Input	User’s Password.
Output	Encrypted - Hashed data.
Precondition	Database must contain all of user’s information and empty columns to store the hashed data.
Post-condition	User’s data are now hashed and safely stored in the database.
Dependencies	The user’s sign up.
Priority	8/10

3.6 Delete

Table 10: deleteAccount

Description	This function is done to make the user able to delete his account.
Action	Check that user exists and let ”Is-Deleted” row is enabled.
Input	None.
Output	Confirmation of successfully deleting the account.
Precondition	User Id is found.
Post-condition	User is deleted from the database.
Dependencies	The user’s log in.
Priority	8/10

3.7 Update

Table 11: updateAccount

Description	This function is done to make the user able to update his account information.
Action	Check that user exists and update his profile information with the new written ones.
Input	New user information.
Output	Confirmation of successfully updating the account.
Precondition	User Id is found.
Post-condition	User's information is updated in the database.
Dependencies	The user's log in.
Priority	8/10

3.8 Classification

Table 12: train-supervised

Description	This function is done to train a supervised data set.
Action	Check that user exists and update his profile information with the new written ones.
Input	Data set file path and methods that would be done(for example word-Ngrams).
Output	Data set information such as the number of labels in the file.
Precondition	Data set is a supervised one.
Post-condition	Data set file training is done.
Dependencies	The existence of supervised data set.
Priority	9/10

Table 13: predict

Description	This function is done to predict the label of a certain input text.
Action	Take a certain input text and figure out the most relevant label to it.
Input	The text that would be classified and number of K.
Output	K numbers of labels that the text would be classified into them and the probabilities of each label.
Precondition	Having a trained supervised data set.
Post-condition	The text is classified to K labels.
Dependencies	train-supervised.
Priority	9/10

3.9 Search

Table 14: searchByUrl

Description	This function is done to retrieve the web content of a certain web page and print it on the screen.
Action	if the input is null return error message else check if the URL exists or not if yes return URL content.
Input	The Website URL
Output	Web Page content of the input url.
Precondition	URL already exists.
Post-condition	Content will be retrieved.
Dependencies	The existence of the input Url.
Priority	7/10

3.10 TwitterApi

Table 15: writeTweet

Description	This function is done to make the user able to create a tweet on his Twitter account through the application.
Action	After getting all of the required keys to start like consumer key, consumer secret, access token and access token secret, the function ”post” will be used to take the input text from the user and apply it to the registered Twitter account to be tweeted.
Input	The input text ”tweet”.
Output	The tweet successfully sent to the account.
Precondition	Twitter Account with required keys exists.
Post-condition	Tweet will be sent.
Dependencies	The existence of account with its keys.
Priority	8/10

Table 16: readTweet

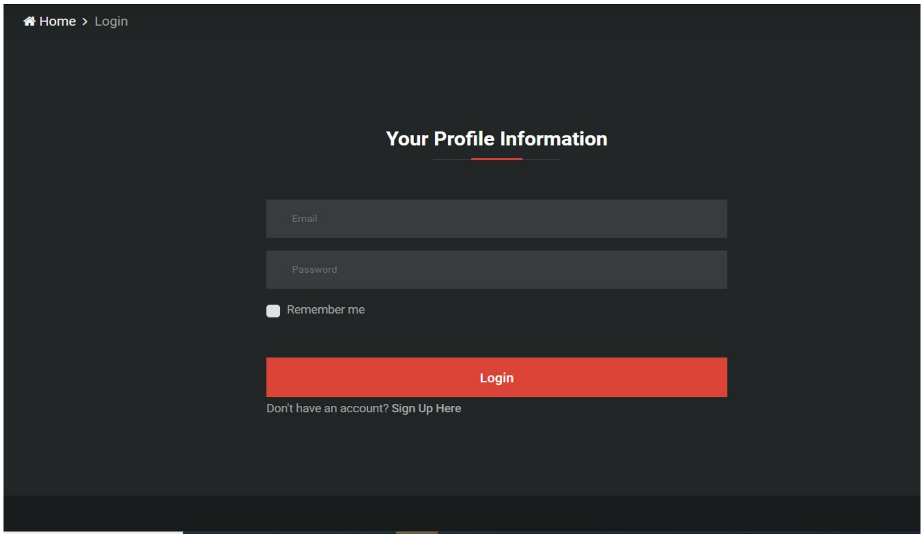
Description	This function is done to make the user able to read tweets from his twitter account directly from the application.
Action	After getting all of the required keys to start like consumer key, consumer secret, access token and access token secret, the function ”get” will be used to access the twitter timeline of this user to get tweets which have a parameter of ”counts” which can be used to determine how many tweets will be retrieved.
Input	None.
Output	Tweets successfully retrieved.
Precondition	Twitter Account with required keys exists.
Post-condition	Tweets will be retrieved.
Dependencies	The existence of account with its keys.
Priority	8/10

4 Interface Requirements

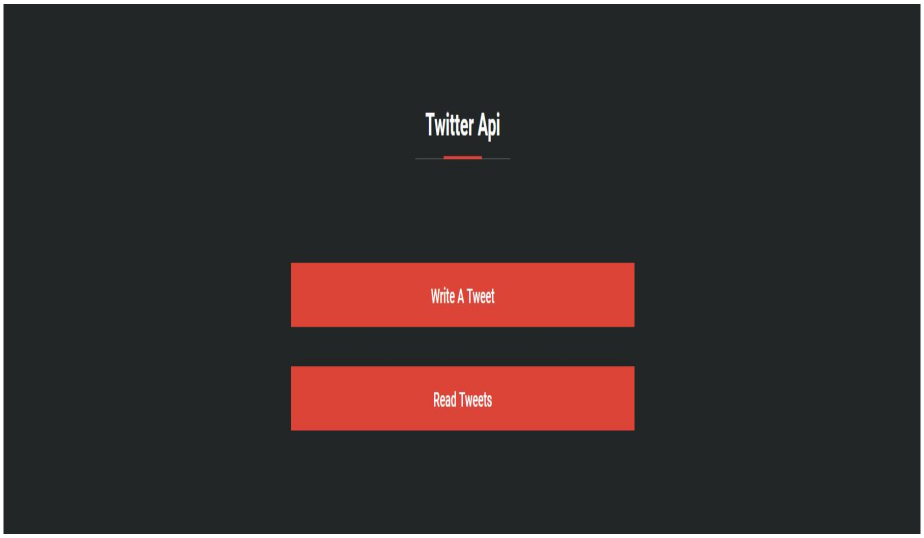
4.1 User Interface

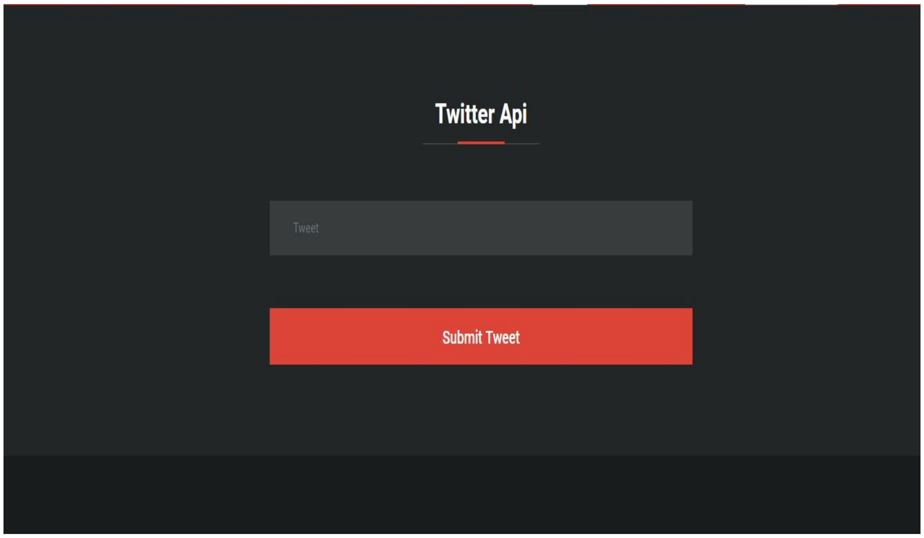
4.1.1 GUI

1. Login Screen

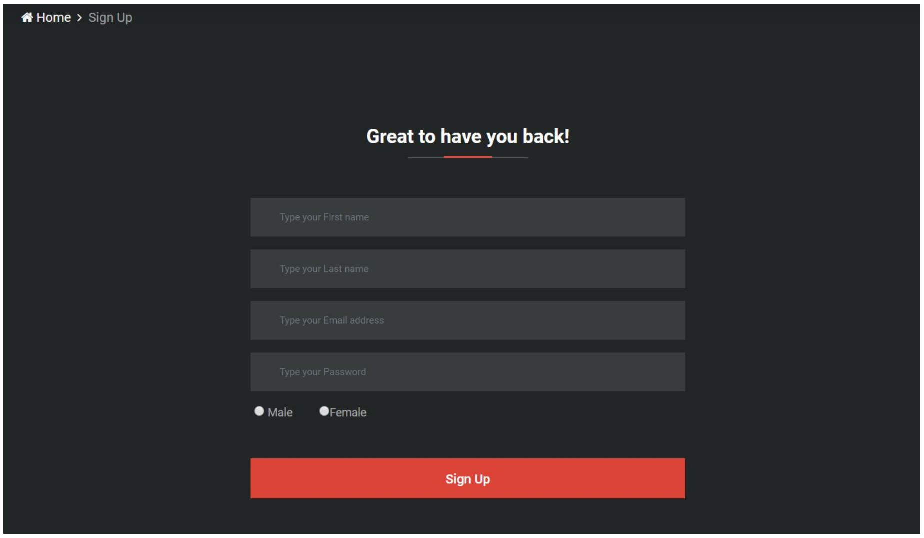


2. Twitter Api Screen

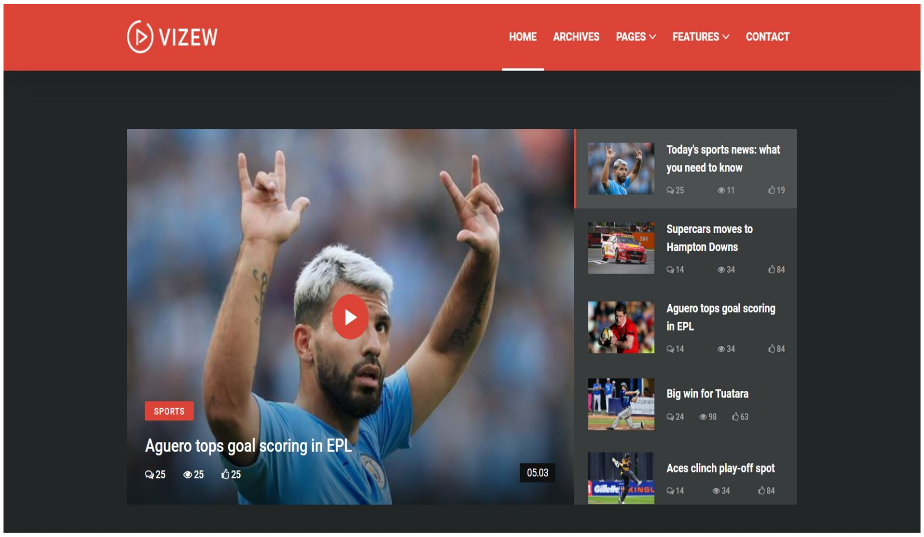




3. Sign Up Screen



4. User - Reader Home Screen



Aguero tops goal scoring in EPL

25 25 25

Today's sports news: what you need to know

25 11 19

Supercars moves to Hampton Downs

14 34 84

Aguero tops goal scoring in EPL

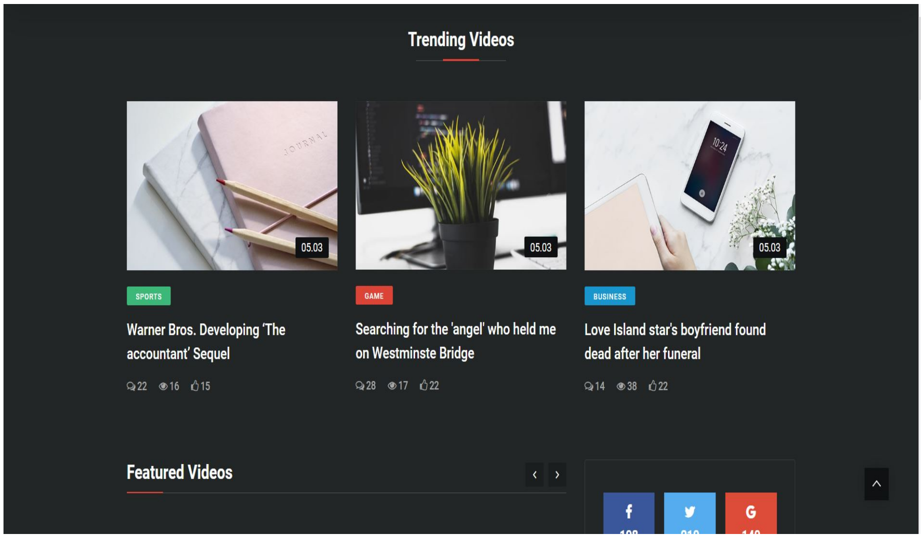
14 34 84

Big win for Tuatara

24 98 63

Aces clinch play-off spot

14 34 84



Warner Bros. Developing 'The accountant' Sequel

22 16 15

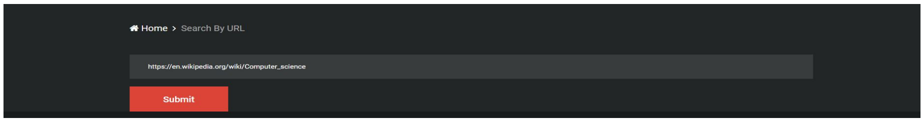
Searching for the 'angel' who held me on Westminste Bridge

28 17 22

Love Island star's boyfriend found dead after her funeral

14 38 22

5. Search By Url Screen



https://en.wikipedia.org/wiki/Computer_science

4.1.2 CLI

N/A

4.1.3 API

1. Facebook FastText API
2. Twitter API
3. Python Libraries
   a) Nltk Libraries
   b) Numpy Libraries
   c) Csv Libraries

4.1.4 Diagnostics or ROM

N/A

4.2 Hardware Interfaces

N/A

4.3 Communication Interfaces

Internet Connection or localhost connection will be needed in the stage Communication interface as it is obligatory to have one connection of them in order to explore the application.

4.4 Software Interfaces

Microsoft Excel is needed so that news data set would be imported to the system.

5 Performance requirements

The system will contain a large number of articles and people's reviews so it needs a strong and fast system that will be well handled and perform the best output in a quick and fast way to show a summarized article and the people's reviews about this article. So the processing stage must never take a lot of time to finish its functionality.

6 Design Constraints

6.1 Standards Compliance

1. 64 or 32-bit operating system.
2. PC with 4.00 GB of RAM (Minimum).

6.2 Hardware Limitations

N/A

7 Non Functional requirements

7.1 Security

User's Id and Password should be encrypted by hashing functions in order to be securely saved in the database.

7.2 Maintainability

Our system should be maintainable enough because it could be improved in the future by other developers, so the design and code should be documented by using different design patterns such as Model View Controller design pattern and Singleton design pattern.

7.3 Portability

The system should be compatible with both Computer and Mobile Devices, so it can be accessed easily and accessible at anytime.

7.4 Usability

The system is easy to use as its functionality isn't so difficult so it won't need a lot of time to be learned.

8 Preliminary Object-Oriented Domain Analysis

8.1 Class diagram:

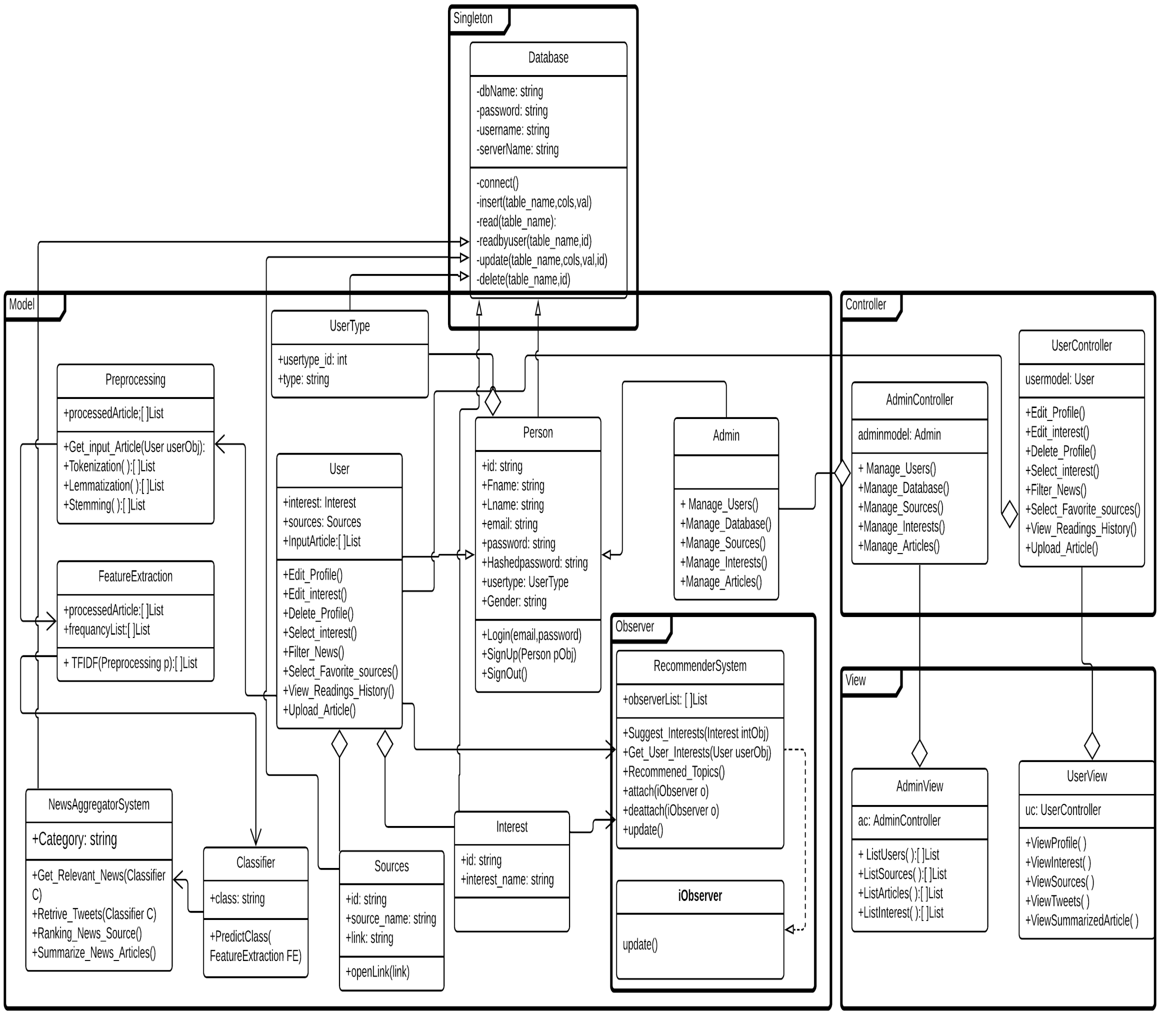


Figure 4: Class diagram

9 Operational Scenarios

9.1 Use Case:

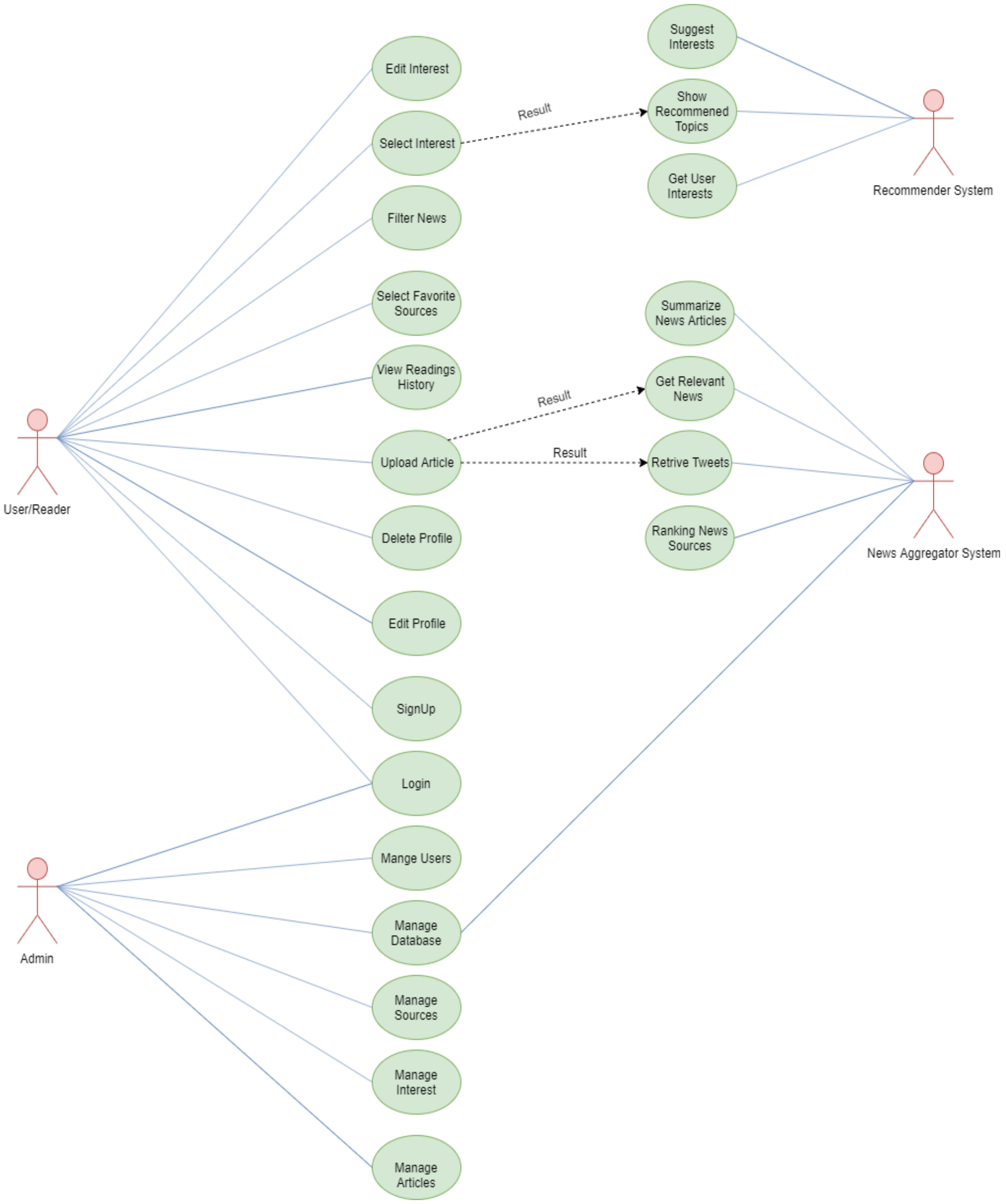


Figure 5: Use Case

According to our system, The main actors would be as the following :

9.1.1 User

The user which is defined as the reader can register an account and edit the profile or delete it.While registering he can select interests and select the favourite news sources.When the user start to read, he can upload an article in order to get the relevant articles from news sources.In addition, he can view his reading history or filter the news search result.

9.1.2 Admin

The admin will be an user that can control the system where he can manage the sources,database,interests,sources and news articles. Admin can add, delete and modify the users, he can also modify and choose the news sources.Articles can be modified by the admin

9.1.3 Recommender System

According to the user, this system can show recommended topics upon user interests. Beside that it can also suggest the user some new interests regardless his chosen interests.

9.1.4 News Aggregation System

This is our main system that can get relevant article and tweets using twitter API according to the user input. It also summarize the output articles. In addition, it rank all the relevant sources to avoid contradiction between news articles.

10 Preliminary Schedule Adjusted

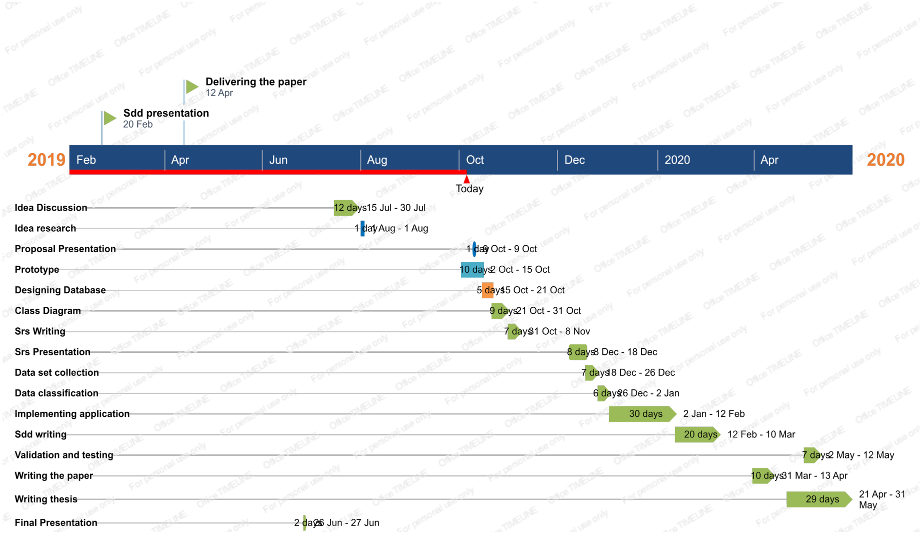


Figure 6: Gantt Chart for time plan

References

[1] Cosmin Grozea, Dumitru-Clementin Cercel, Cristian Onose, and Stefan Trausan-Matu. Atlas: News aggregation service. In *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, pages 1–6. IEEE, 2017.

[2] Kimberley A Isbell. The rise of the news aggregator: Legal implications and best practices. *Berkman Center Research Publication*, (2010-10), 2010.

[3] Jaka E Sembodo, Erwin B Setiawan, and ZK Abdurahman Baizal. The improvement of indonesian news curator classification in twitter. In *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, pages 1–7. IEEE, 2017.

[4] K Sundaramoorthy, R Durga, and S Nagadarshini. Newsone—an aggregation system for news using web scraping method. In *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, pages 136–140. IEEE, 2017.