Newspaper Article Summarizer

A Project

Presented to the

University of Dhaka, Bangladesh

In Partial Fulfillment

Of the Requirements for the Degree

Master in Information Technology (MIT)

By

Name: Md. Siam Ansary

Roll: 201103

November 2021

SIGNATURE PAGE

PROJECT: Newspaper Article Summarizer

AUTHOR: Md. Siam Ansary

DATE SUBMITTED: 10 November 2021

SUPERVISED BY: Dr. B M Mainul Hossain
Associate Professor
Institute of Information Technology

SUPERVISOR'S APPROVAL: 10-11-21

ACKNOWLEDGEMENTS

I am very grateful and thankful to my kind supervisor, Dr. B M Mainul Hossain. Without his sincere guidance and help, I would not be able to complete the project work. It has been a privilege to work under his guidance.

I am also very grateful to all faculty members of Institute of Information Technology. I have gained much knowledge from them which helped a lot in this work. I am thankful for their support and cooperation.

Profound thanks are due to my family and classmates for their inspiration and cooperation.

Lastly, I am indebted to everyone who supported me in any respect during the completion of the project work.

ABSTRACT

Text summarization produces a concise amount of data out of document which may represent the core information contained in the document. Summaries should have all the important information and no crucial data should be left out. In recent times much research is being done in automatic text summarization. Having an overview of the state-of-the-art for automatic text summarization, it can be said that, this area is still very much open and there are many scopes for improvements. In this project, a system has been used to generate summary from newspaper article that uses combination of features to rank sentences of the article. Graph based approach with minimal text sentiment analysis and the presence of category wise key phrases have been used. It has been observed that using graph, the interconnection between sentences and thus most significant sentence can be identified. Sentences with the sentiment polarity of the overall document sentiment polarity can convey important information. Category wise key phrases can emphasize a sentence that can have salient data of a text. Combination of multiple features can lead to the identification process of the significant sentences with which good summary can be formed. Hence, in this project, a combinational approach is used to generate summary. After the performance analysis of the approach, a simple application has been developed where a user can fetch a newspaper article text providing the URL of the article, summarize and can store the summary.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1.Objectives

A summary gives the overall sense of a document. Hence in different works, instead of going through the full document, from summary important information can be gained which makes the overall procedure more comfortable and less resources are needed in the long run. All the major topics of the document should be in the summary. If the summary contains sentences from all the major topics present in the document that has a better chance of giving better perspective of the document. It is typical of human summaries, too. When humans summarize a document, they include most of the topics present in the document in their summaries. Text Summarization approaches are mainly of two categories. They are extractive summarization and abstractive summarization. The extractive summarization is the one where the exact sentences present in the document are used as summaries. The extractive summarization is simpler and is the general practice among the automatic text summarization researchers at the present time. Extractive summarization process involves giving scores to sentences using some method and then using the sentences that achieve highest scores as summaries. Although this kind of summary does its job in conveying the essential information, it may not be necessarily smooth or fluent. Sometimes there can be almost no connection between adjacent sentences in the summary, resulting in the text lacking in readability. The abstractive summarization is the process in which the abstract of the document is created. The abstract can contain the words and phrases not present in the original document. The abstractive summarization procedure is a very complicated process as the semantics of sentences has to be dealt with. Several other factors such as word sense, grammatical structure has to be taken into consideration before creating a useful abstractive summary.

The summary generation approach of this project is extractive, i.e., the summaries contain exact sentences present in the document. Residing on the idea that a good summary must cover all major information present in the document, the first step towards text summarization is naturally to identify the important parts of texts. To accomplish this task, it is a must to extract the sentences which convey the significant information. Hence, sentences need to be scored which can be done in many approaches. A sentence that is hugely connected to other sentences that may convey the culmination of information is important. Some textual portions may carry notable emotional data. Therefore, sentiment analysis may play a vital part in identifying sentences that ought to be included in summary. Key-phrases in different categories emphasize on material of sentence which may lead to be included in summary. The generated summaries needs to be evaluated to determine the effectiveness and efficiency of approaches.

The objective of the project is to explore a combinational extractive approach for text summarization of newspaper articles and to develop an application so that a user can easily summarize articles.

1.2. Motivations

The field of text summarization is developing day by day. There have been many attempts to construct a summarizer that produces summary that is close to human's work. Most of the works have been done so far are mainly on extractive approach though research on abstractive process is also being done. Different approaches considers different features for improving the results. Sentence features, graph approaches, machine learning are few of the approaches. Some works might even have contradicting motivations. Elbarougy et al. [1] have worked on Arabic text summarization with extractive approach. Since the Arabic language has a complex morphological structure, it can be very difficult to do automatic summarization. The researchers have used modified PageRank algorithm where the initial score of each node is the number of nouns in sentence. They have opinionated that nouns in the sentence increase its importance and more nouns means more information in a sentence. For the evaluation purpose, the Essex Arabic Summaries Corpus has been used as a standard corpus as the corpus contains 153 documents, and each document has 5 summaries, with a total of 765 Arabic human-made summaries. Manjari [2] has worked on multi document extractive summarization of Telegu language using TextRank algorithm. For the similarity matrix, cosine similarity has been used. Indic NLP Library from Python Package Index(PyPI) [3] has been used for various processing. Sharma and Jain [4] have done research on sentiment analysis and summarization on the student feedback system of Institute of Engineering and Technology, Devi Ahilya Vishwavidhyalaya, Indore of India. For the summarization task, TextRank has been employed. Mane et al. [5] have done extractive summarization using health posts from social networking sites. They extracted keywords and categorized them using Unified Medical Language System. Then, sentences have been clustered based on categorized keywords and sentence scoring has been done based on presence of keywords. Since presence of topic key-phrases has been used for sentence scoring in different experiments, Aditi et al. [6] have looked into which technique can be more suitable for determining topic key phrases. They have observed that, between Word Probability and TF-IDF, TF-IDF is better.

Since there are many experiments being done on the topic of summarization, in this project, a hybrid approach has been explored and a simple tool employing the approach has been developed.

1.3. Scope

In the project, a combinational approach has been experimented with for sentence scoring to do extractive summarization of newspaper articles. Presence of article category based keywords, sentiment analysis and TextRank algorithm have been used. Five newspaper categories have been considered, they are business, entertainment, politics, sport and tech. Through sentiment analysis, it has been observed that how a sentence’s sentiment polarity can be used to identify more significant information.

After the performance analysis, a tool has been developed where a user has to input the URL of an online newspaper article to fetch the article. Since different websites can have different structures, newspaper article from all websites may not be successfully fetched. If the article is successfully fetched, the hybrid approach is employed to score the sentences and with top ranking forty percent sentences, summary is formed. A user can have the option to save the summary in the local database or a text file.

Since public articles are to be fetched for the summarization work, no security aspects have been considered in the work.

1.4. Description

In this project, at first, research work has been done and then development has been done.

In the research work, it has been observed how a combinational approach can be used to do summarization of newspaper articles. BBC News Summary dataset from Kaggle [7] platform has been used for research work. The dataset has newspaper articles of five categories and their summaries. With approximately fifty percent articles of the five categories, five keywords lists are prepared. On the rest of the articles, the hybrid approach has been applied and the performance has been observed.

In the combinational approach, three methods are used.

- TextRank algorithm has been used to score all the sentences. It is language independent and unsupervised, hence, no training is needed.
- For each sentence, keyword extraction has been done. Then, it has been checked how many of a sentence’s keywords are in the prepared keyword list of the article’s category. Then, the number is normalized with sentence length.
- Through sentiment analysis, the polarity of whole document is at first determined. Then, sentiment analysis is done on each sentence. If a sentence’s sentiment polarity is same as the whole document’s polarity, then the sentence is prioritized for summarization. It has been checked, if neutral sentences are only prioritized rather than prioritizing sentences whose polarity matches with the overall document’s sentiment polarity. It has been found out, better performance can be achieved if the sentences whose sentiment polarity matches document polarity are considered rather than considering only neutral ones.

After the experimental research phase, an application has been developed which is very simple. A user has to input his/her username, URL of a newspaper and the category of the article. If the article is successfully fetched, then summarization is done. The user can have the options to save the generated summary is a text file and save the summary is local database. The user can check the stored records from the database, save a stored generated summary in text file and delete a record.

Chapter 2

Methodology

2.1 Requirements

Based on the hybrid summarization approach, an application has been developed. Requirements for the application have been identified as below.

- Proper input fields: With the application, a user can fetch a newspaper article in order to do summarization. S/he will have proper input fields to put her/his username, URL of the newspaper and category of the article.
- Retrieval of newspaper article text with URL: If the user inputs URL of a newspaper article from a well-structured website, the application must be able to fetch the article text.
- Generated summary saving: The user will be able to save the generated summary in a text file and also in the local database of the application.
- Use of local database for stored summaries: If a summary has been saved in the database of the application, user will be able to observe the information associated with the record, save in text file and delete a record.

2.2 Technology

For experimental research and development works, Python programming language and its different packages have been used. Different packages of Python can be found on PyPI [3].

2.3 Summarization Approach

The summarization process is a combination of three techniques. They are

- Interconnection of sentences using TextRank algorithm
- Prioritization of sentences of sentiment polarity similar to overall document sentiment
- Primacy of sentences with category-wise keywords

Before the summarization, different keywords lists have been prepared for each category of the categories Business, Entertainment, Politics, Sports and Tech using approximately fifty percent data of the BBC Newspaper Summary dataset from Kaggle [7]. The key-phrase extraction is done by the approach proposed by Rose et al. [8]. They have proposed Rapid Automatic Keyword Extraction which is unsupervised, domain independent and language independent and extracts keyword from individual documents. It is very simple yet has computational efficiency.

In the summarization approach at first a newspaper article goes through pre-processing as stop-words are removed. After stop-words removal, the sentences are scored in three methods and the scores are combined

```
Algorithm 1: Steps for key phrase extraction
Step 1   Start
Step 2   Input document
Step 3   Split the document into an array of words
Step 4   Break array of words at word delimiters
Step 5   Split the words into sequences of contiguous words
Step 6   Break each sequence at a stop-word. Consider each sequence as a candidate keyword
Step 7   Calculate the "score" of each individual word in the list of candidate keywords using the metric $degree(word)/frequency(word)$
Step 8   For each candidate keyword, add the word scores of its constituent words to find the candidate keyword score
Step 9   End
```

In the summarization approach at first a newspaper article goes through pre-processing as stop-words are removed. After stop-words removal, the sentences are scored in three methods and the scores are combined.

For the interconnection based scoring, the text is represented as a graph and TextRank algorithm is applied. Cosine similarity metric used for calculating similarity between two sentences

Through sentiment analysis, the overall document's sentiment polarity is determined. Then, sentiment polarity of each sentence is determined. If a sentence's sentiment polarity is same as the document's polarity, it is considered as a candidate summary sentence.

Since the category of each newspaper article is known, it is identified how many key-phrases there are in a sentence from the article's category's key-phrase list that has been prepared beforehand. The count is normalized with sentence length, since longer sentences can have more key-phrases.

Hence, for each sentence, s_j

$$score\ (s_j) = \sum_{i=1}^{3} criterion_i score(s_j) \qquad \text{(eqn 1)}$$

$$criterion_1 score(s_j) \ = TextRank\ score(s_j) \qquad \text{(eqn 2)}$$

$$criterion_2 score(s_j) = 1\ if\ sentiment\ polarity\ of\ (s_j) \text{ is same as document} \qquad \text{(eqn 3)}$$

$$criterion_2 score(s_j) = 0\ if\ sentiment\ polarity\ of\ (s_j) \text{ not same as document} \qquad \text{(eqn 4)}$$

$$criterion_3 score(s_j) \ = \frac{category\ keywords\ count\ in\ sentence_j}{length\ of\ sentence_j} \qquad \text{(eqn 5)}$$

After all the sentences are scored, the top scoring forty percent sentences are used to form summary.

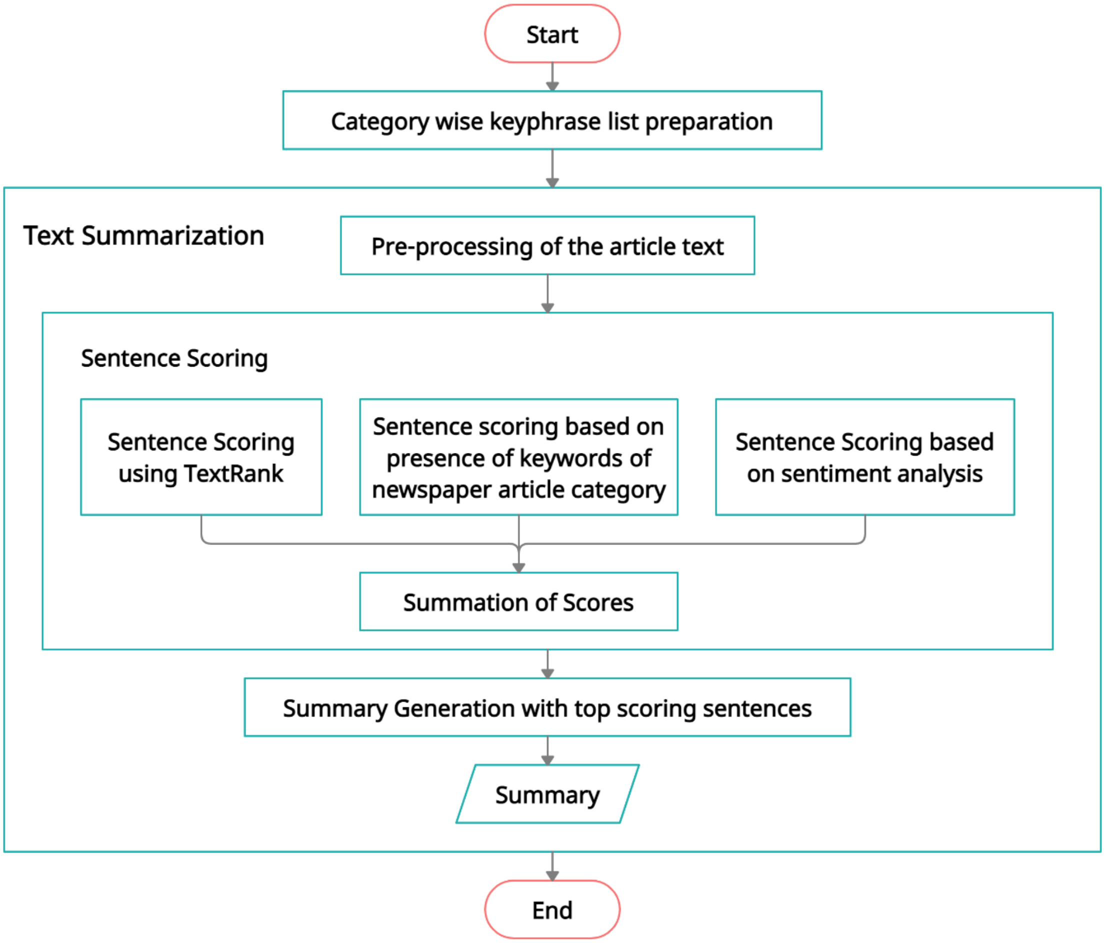

Start
Category wise keyphrase list preparation
Text Summarization
Pre-processing of the article text
Sentence Scoring
Sentence Scoring using TextRank
Sentence scoring based on presence of keywords of newspaper article category
Sentence Scoring based on sentiment analysis
Summation of Scores
Summary Generation with top scoring sentences
Summary
End

Figure 1. Steps of the combinational summarization approach

Based on this experimental summarization approach, the summarization application is developed.

Chapter 3

Implementation

3.1 Implementation Details

For the experimental summarization approach, TextRank algorithm, key phrase extraction and sentiment analysis have been done. Also, category wise key phrases lists have been prepared. While implementing these, different Python packages have been used. For different language processing works, Natural Language Tool Kit (NLTK) and Scikit-learn have been used. For sentiment analysis, TextBlob package; for key phrase extraction RAKE package, for TextRank implementation, NetworkX package, for saving the phrase lists, Pickle package have been used.

For the application development, based on the experimental approach, Tkinter package has been used for Graphical user interface related works. Generated summaries can be saved using the SQLite3 module. When the user provides an URL, the Newspaper package is used to fetch the newspaper article.

When the user uses the application, s/he has to provide three pieces of information. They are

- Username
- URL of the newspaper article
- Category of the article

Then the user can fetch the article. Once the article is fetched, the user can opt to summarize the text. After summarization is done, the summary text will be displayed. The user can save the summary is database and a text file.
The user can check the recorded summaries in the database. All the records are to be displayed. The user can do the following on the recorded entries.

- Search the recorded entries by username
- Select a particular record
- Save the particular record’s summary is text file
- Delete the particular selected record

In the implementation, security aspect of any kind has not been considered as one has to use publicly available newspaper URL for summarization task in the application.

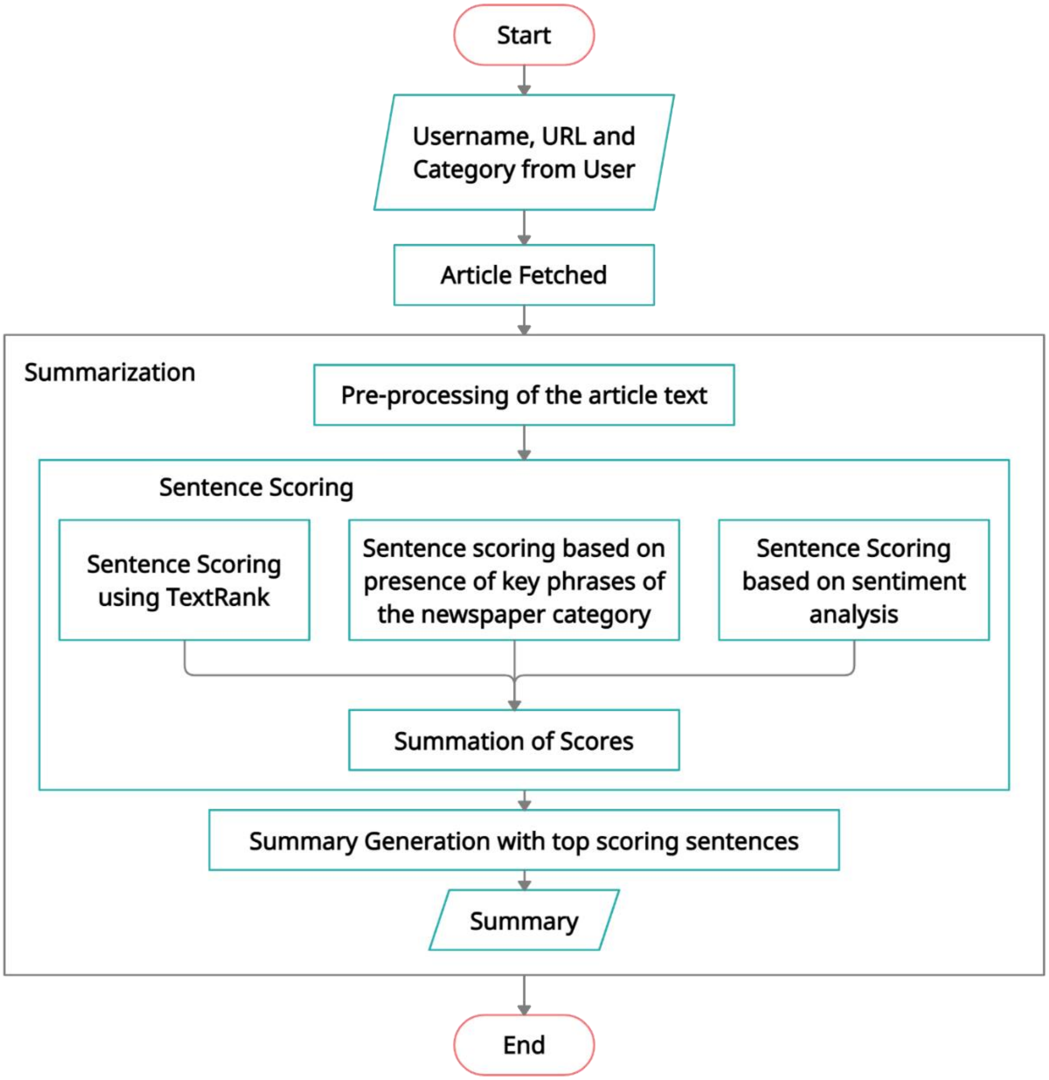

Start
Username, URL and Category from User
Article Fetched
Summarization
Pre-processing of the article text
Sentence Scoring
Sentence Scoring using TextRank
Sentence scoring based on presence of key phrases of the newspaper category
Sentence Scoring based on sentiment analysis
Summation of Scores
Summary Generation with top scoring sentences
Summary
End

Figure 2. Summarization steps in the developed application

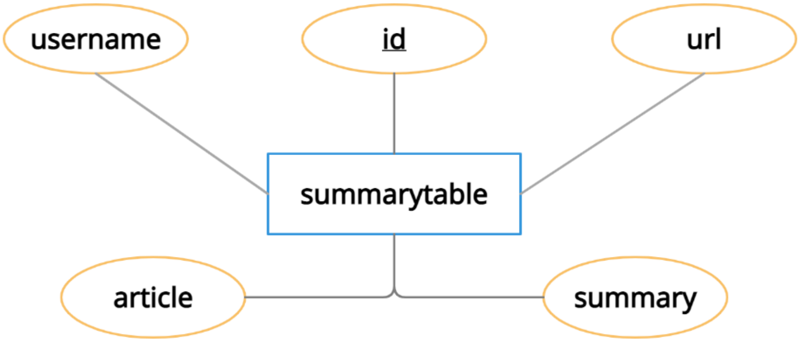
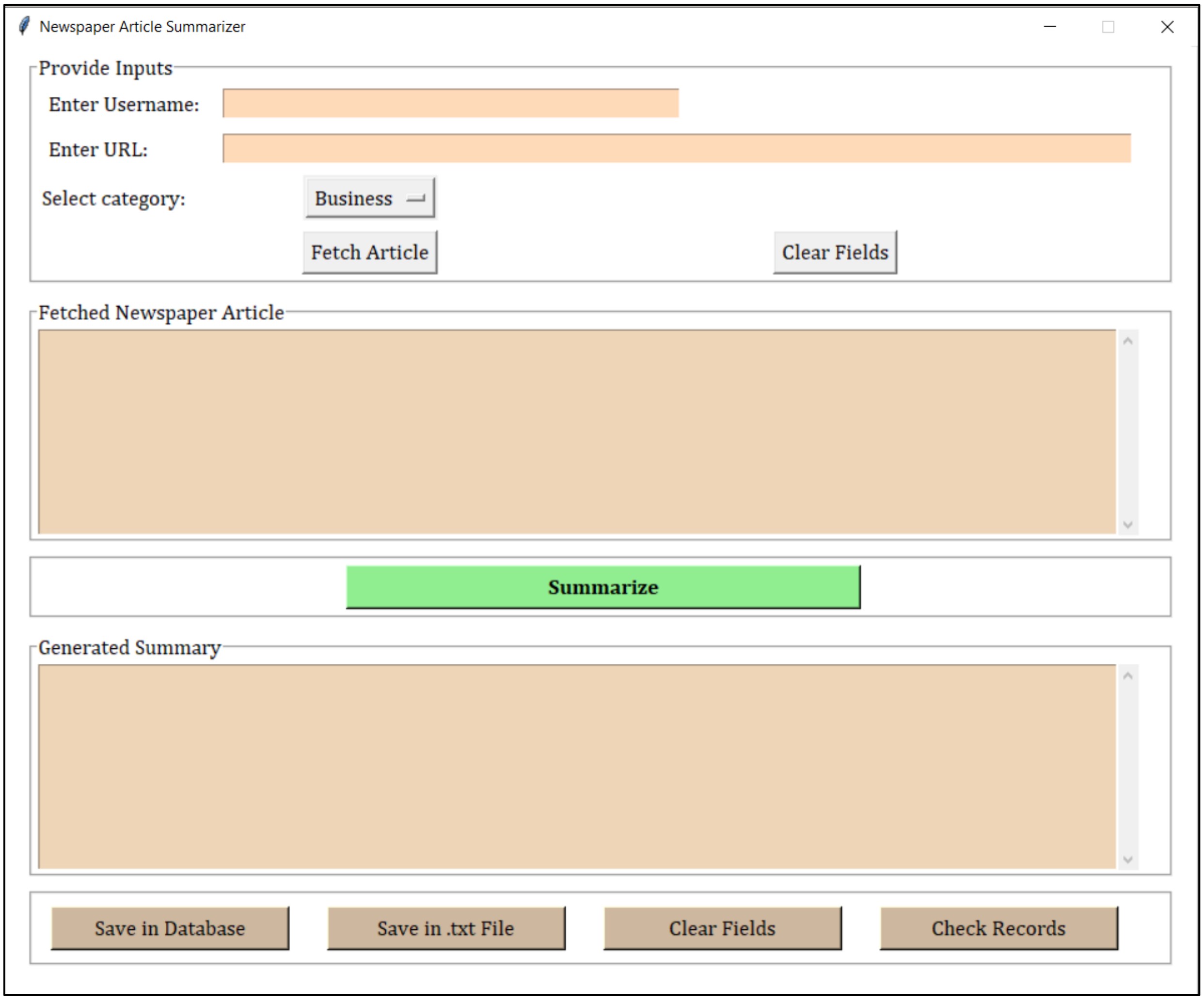


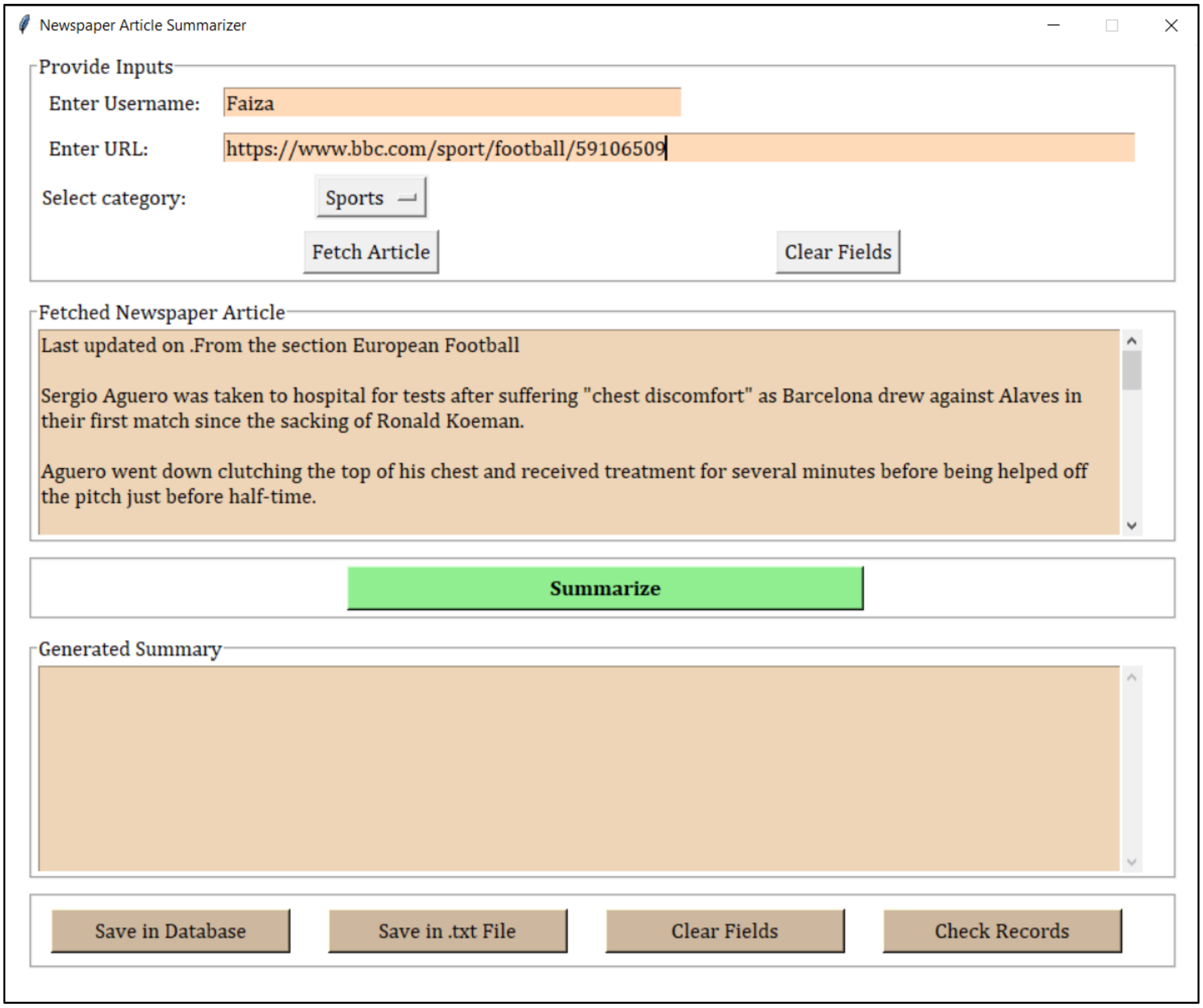
Figure 3. ERD of the database table of the application

The ERD diagram of the database table where summaries are stored is shown in Figure 3. In the figures 4 to 7, some aspects of the application can be observed.



Select category: Business

Figure 4. GUI for user to provide inputs and for summary display with other options



Newspaper Article Summarizer

Provide Inputs

Enter Username: Faiza

Enter URL: https://www.bbc.com/sport/football/59106509

Select category: Sports

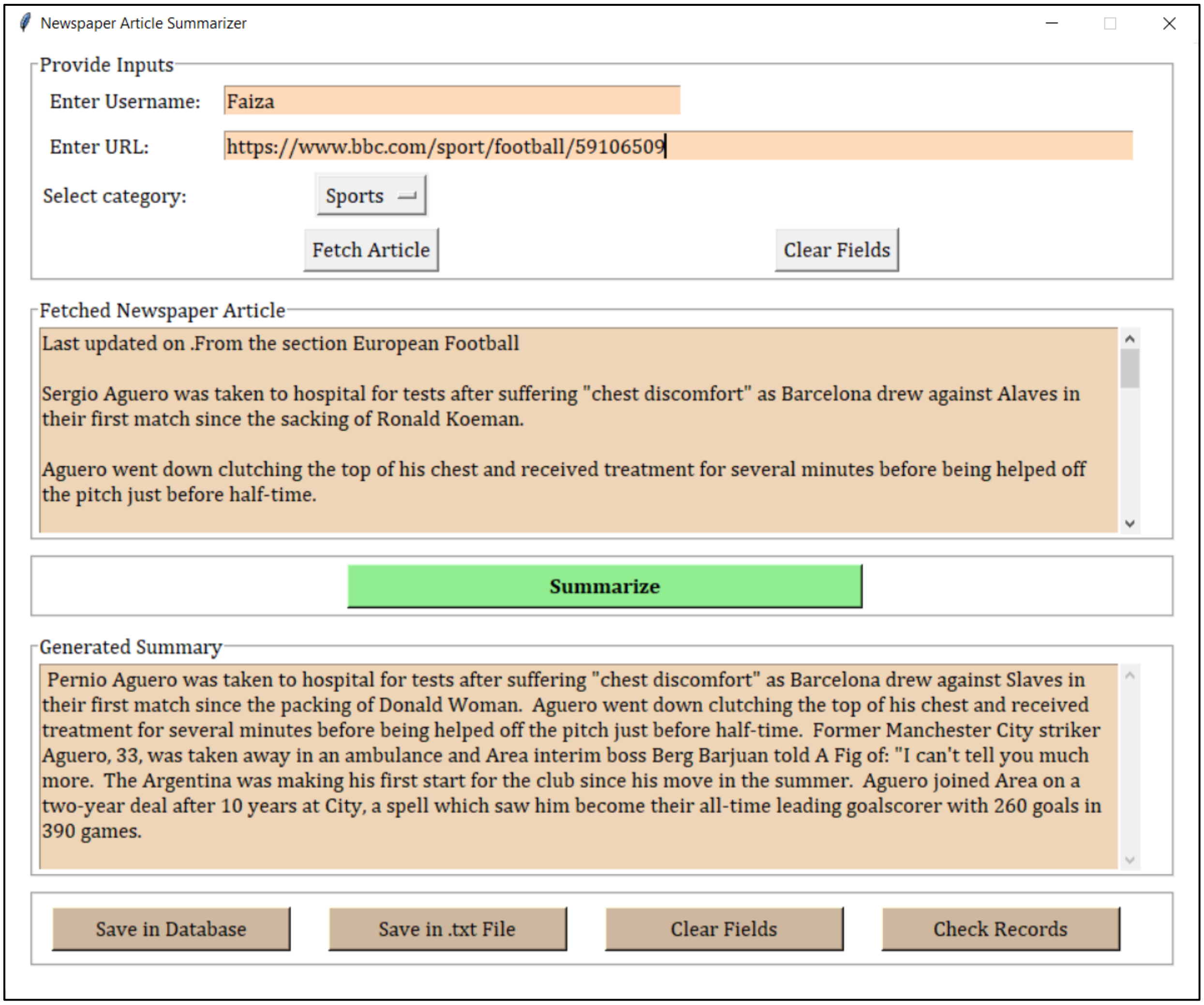
Fetched Newspaper Article

Last updated on .From the section European Football

Sergio Aguero was taken to hospital for tests after suffering "chest discomfort" as Barcelona drew against Alaves in their first match since the sacking of Ronald Koeman.

Aguero went down clutching the top of his chest and received treatment for several minutes before being helped off the pitch just before half-time.

Figure 5. User provides inputs and an article is fetched



Enter Username: Faiza

Enter URL: https://www.bbc.com/sport/football/59106509

Select category: Sports

Fetched Newspaper Article

Last updated on .From the section European Football

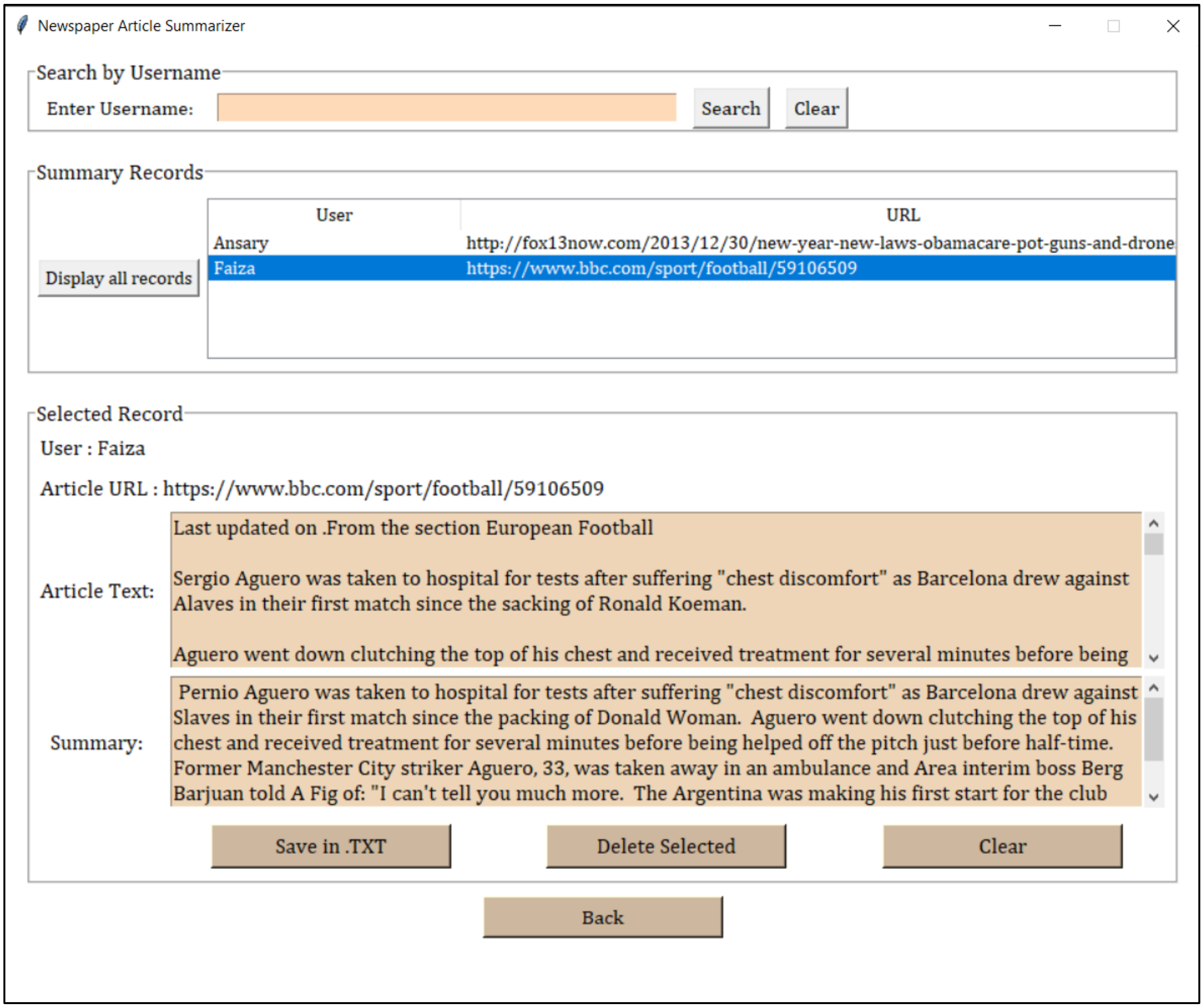
Sergio Aguero was taken to hospital for tests after suffering "chest discomfort" as Barcelona drew against Alaves in their first match since the sacking of Ronald Koeman.

Aguero went down clutching the top of his chest and received treatment for several minutes before being helped off the pitch just before half-time.

Generated Summary

Pernio Aguero was taken to hospital for tests after suffering "chest discomfort" as Barcelona drew against Slaves in their first match since the packing of Donald Woman. Aguero went down clutching the top of his chest and received treatment for several minutes before being helped off the pitch just before half-time. Former Manchester City striker Aguero, 33, was taken away in an ambulance and Area interim boss Berg Barjuan told A Fig of: "I can't tell you much more. The Argentina was making his first start for the club since his move in the summer. Aguero joined Area on a two-year deal after 10 years at City, a spell which saw him become their all-time leading goalscorer with 260 goals in 390 games.

Figure 6. Generated summary is displayed



User	URL
Ansary	http://fox13now.com/2013/12/30/new-year-new-laws-obamacare-pot-guns-and-drone
Faiza	https://www.bbc.com/sport/football/59106509

User : Faiza

Article URL : https://www.bbc.com/sport/football/59106509

Article Text: Last updated on .From the section European Football

Sergio Aguero was taken to hospital for tests after suffering "chest discomfort" as Barcelona drew against Alaves in their first match since the sacking of Ronald Koeman.

Aguero went down clutching the top of his chest and received treatment for several minutes before being

Summary: Pernio Aguero was taken to hospital for tests after suffering "chest discomfort" as Barcelona drew against Slaves in their first match since the packing of Donald Woman. Aguero went down clutching the top of his chest and received treatment for several minutes before being helped off the pitch just before half-time. Former Manchester City striker Aguero, 33, was taken away in an ambulance and Area interim boss Berg Barjuan told A Fig of: "I can't tell you much more. The Argentina was making his first start for the club

Figure 7. Possible operations from recorded entries

Chapter 4

Results

4.1 Performance of different public sentiment analysis packages

There a few publicly available sentiment analysis (SA) packages in Python such as TextBlob, VADER, Sentifish. Through experimenting with public datasets from Kaggle [7], it has been found out that in regards to accuracy and time, TextBlob is the most promising one.

The publicly available datasets are

- Sentiment Analysis for Financial News (SAFN)
- Twitter and Reddit Sentiment Analysis

Table 1. Performance of SA packages on SAFN dataset

Package	Accuracy	Total Time (sec)
Sentifish	0.594	44.609
TextBlob	0.491	1.619
VADER	0.543	0.639

Table 2. Performance of SA packages on Reddit Data dataset

Package	Accuracy	Total Time (sec)
Sentifish	0.353	510.682
TextBlob	0.999	11.142
VADER	0.644	18.716

Table 3. Performance of SA packages on Twitter Data dataset

Package	Accuracy	Total Time (sec)
Sentifish	0.339	1728.557
TextBlob	0.998	39.139
VADER	0.569	26.944

4.2 Performance analysis of experimental summarization approach

In chapter 2, it has been discussed how sentence scoring is done for the experimental text summarization approach. For evaluation, ROUGE measure has been used.

$$ROUGE\ Recall = \frac{number\ of\ overlapping\ words}{number\ of\ words\ in\ reference\ summary} \qquad \text{(eqn 6)}$$

$$ROUGE\ Precision = \frac{number\ of\ overlapping\ words}{number\ of\ words\ in\ the\ generated\ summary} \qquad \text{(eqn 7)}$$

Chin-Yew Lin [9] proposed ROUGE measure that stands for Recall-Oriented Understudy for Gisting Evaluation. A public implementation is available in PyPI [3].

Table 4. Performance of the experimental summarization approach

Article Category	ROUGE 1 Precision	ROUGE 1 Recall	ROUGE 1 F Score
Business	0.633	0.536	0.47
Entertainment	0.577	0.506	0.455
Politics	0.604	0.546	0.501
Sport	0.564	0.491	0.441
Tech	0.621	0.558	0.51
Average	0.5998	0.5274	0.4754

Now, in the project work, as one of the criteria, through sentiment analysis, the sentences whose sentiment polarity matches with the document sentiment polarity, are prioritized. But, if the experimental system was to be modified in the way that neutral sentences were to be prioritized as a criterion, the performance would have downgraded as shown in the Table 5.

Table 5. Performance of the modified experimental summarization approach

Article Category	Rouge 1 Precision	Rouge 1 Recall	Rouge 1 F Score
Business	0.606	0.482	0.404
Entertainment	0.54	0.439	0.375
Politics	0.592	0.489	0.422
Sport	0.56	0.459	0.393
Tech	0.573	0.468	0.399
Average	0.5742	0.4674	0.3986

Chapter 5

Conclusion

5.1 Conclusion

The experimental approach in the project work is promising but there can be more improvements. There are several lacking and difficulties for implementation processes and different future work can add more progresses.

5.1.1 Difficulties

i. Natural language processing can be very computationally expensive. Hence, lots of resources are necessary. As a result, the summarization phase in the application can be very slow.

ii. Lots of implementations depend on other research topics and public resources. As a result, there can be lacking in implementation.

iii. In text summarization, may research problems are still unexplored. As a result, automatically generated summaries may not be close to human written summaries yet.

5.1.2 Future Works

i. In future, more sentence ranking approaches may be incorporated and performance analysis can be observed.

ii. The experimental approach can be tested on more datasets.

iii. The experiments can be done with languages other than English.

iv. The summarization application can be made more aesthetic and more features can be included.

Appendix A

Source code of the Project

The source codes related to the project work has been kept in a GitHub repository. The link of the GitHub repository is https://github.com/MdSiamAnsary/Newspaper-Article-Summarizer

References

1. Reda Elbarougy, Gamal Behery, and Akram El Khatib, “Extractive arabic text summarization using modified PageRank algorithm,” Egyptian Informatics Journal, vol 21, issue no. 2, pp. 73-81, 2020.
2. K. Usha Manjari, “Extractive Summarization of Telugu Documents using TextRank Algorithm,” in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), Palladam, India, 7- 9 October 2020.
3. Python Package Index (PyPI), Online: https://pypi.org/, Last Accessed: 9 November 2021
4. Neeraj Sharma, and Vaibhav Jain, “Evaluation and Summarization of Student Feedback Using Sentiment Analysis,” in International Conference on Advanced Machine Learning Technologies and Applications, Jaipur, India, 13-15 February 2020.
5. Vinod L. Mane, Suja S. Panicker, and Vidya B. Patil, “Summarization and sentiment analysis from user health posts,” in 2015 International Conference on Pervasive Computing, Pune, India, 8-10 January 2015.
6. Aditi, Shikha Shandilya, Nidhi Bansal, and Shuchi Mala, “An Evaluation of Word Frequency Techniques for Text Summarization Using Sentiment Analysis Approach,” in 2020 10th International Conference on Cloud Computing, Data Science and Engineering (Confluence), Noida, India, 29- 31 January 2020.
7. Kaggle, Online: https://www.kaggle.com/datasets , Last Accessed: 9 November 2021
8. S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic Keyword Extraction from Individual Documents,” Text Mining: Applications and Theory, pp. 1-20, 2010.
9. Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of summaries”, in proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, January, 2004.