Newspaper Article Summarization using Combinational Method

***Abstract*—Text summarization produces a concise amount of data out of document which may represent the core information contained in the document. In this paper, a combinational approach has been used to generate summary from newspaper article that uses a number 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. Sentences with the sentiment polarity of the overall document sentiment polarity can convey important information. Combination of multiple features can lead to the identification process of the significant sentences with which good summary can be formed. After the performance analysis of the approach, it has been observed that the technique can generate good summaries that are close to human written ones.**

***Index Terms*—Text Summarization , Newspaper Article Summarization , Natural Language Processing**

I. Introduction

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 process is comparatively simpler and is the general practice among the automatic text summarization researchers at the present time. This 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 paper 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 work has been 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.

II. Related Works

The field of text summarization is developing day by day. There have been many attempts to construct a summarization technique 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 keyphrases 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.

Tsai et al. [7] discussed text summarization signifying on importance of text content and sentiment analysis. They worked with online hotel reviews. According to them, it is important to remove redundant or insignificant texts from a document before doing summarization, otherwise noisy data can remain in generated summaries. They valued features of the texts, applied sentiment analysis and through clustering picked the summary sentences.

Study of Meena et al. [8] indicate that with hybrid models, text frequency pick summary sentences with great accuracy.

Nathonghor and Wichadakul [9] discussed text summarization for travel news in Thai language. From their study, it has been evident that if keyphrases are scored correctly, the feature can be used for summary generation.

Naidu et al. [10] discussed summarization of newspaper articles. They prepared a list of salient keywords through extraction and identified summary sentences with significance substance using a parts of speech tagger and presence of keyphrases from the prepared list.

III. Methodology

A. Implementation Approach

In our work, it has been observed how a combinational approach can be used to do summarization of newspaper articles. 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.

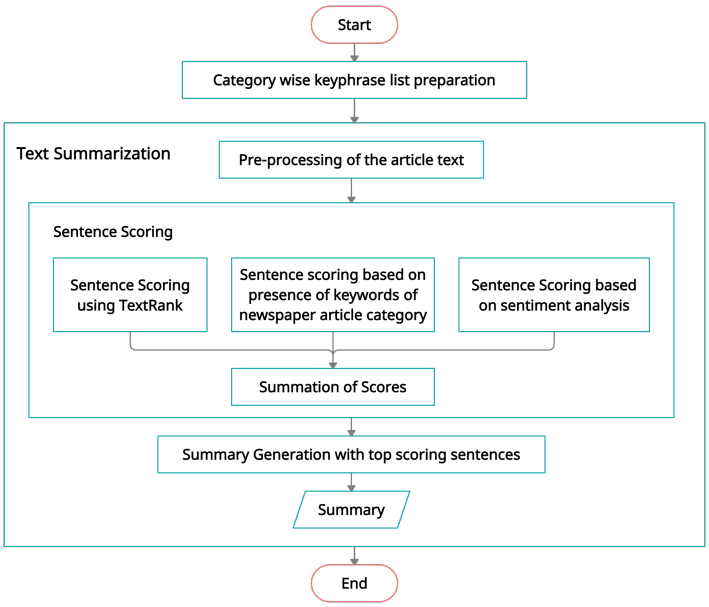

Fig. 1. Flowchart of steps for summarization

We have focused on interconnection among sentences, sentiment analysis and presence of keyphrases. For sentiment analysis based scoring, we have tried two approaches. In *Approach 01*, we have put significance on the sentences whose sentiment polarity is same as the overall sentiment polarity of the document and in *Approach 02*, we have put importance on neutral sentences.

1) Approach 01: For the interconnection based scoring, the text is represented as a graph and TextRank algorithm is applied. GloVe [14] has been used for vector representation of the texts and Cosine similarity metric has been used for calculating similarity between two sentences.

$$t_1^1(s) = TextRank\_Score(s) \tag{1}$$

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.

TABLE I
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 II
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

$$t_2^1(s) = \begin{cases} 1 \text{ if sentence s has same sentiment polarity as document} \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

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.

$$t_3^1(s) = Category\_Keyphrase\_Count(s) \div length(s) \quad (3)$$

$$score^1(s) = t_1^1(s) + t_2^1(s) + t_3^1(s) \quad (4)$$

After all the sentences are scored, the top scoring forty percent sentences are used to form summary. Equations (1) to (4) outline the *Approach 01* technique.

2) Approach 02: There is very little but important difference between *Approach 01* and *Approach 02*. In *Approach 02*, for Sentiment Analysis based scoring, neutral sentences have been considered as candidate summary sentences. The rest of the process has been similar to *Approach 01*.

$$t_2^2(s) = \begin{cases} 1 \text{ if sentence s is neutral} \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

$$score^2(s) = t_1^1(s) + t_2^2(s) + t_3^1(s) \quad (6)$$

Equations 1, 5, 3 and 6 recapitulate the technique of *Approach 02*.

B. Dataset

BBC News Summary dataset from Kaggle [11] 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.

TABLE III
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

C. Key Phrase Extraction

The key phrase extraction is done by the approach proposed by Rose et al. [12]. 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.

D. Different Existing Sentiment Analyzers

There a few publicly available sentiment analysis (SA) packages in Python such as TextBlob, VADER, Sentifish. Through experimenting with public datasets Sentiment Analysis for Financial News (SAFN) and Twitter and Reddit Sentiment Analysis from Kaggle [11], it has been found out that in regards to accuracy and time, TextBlob is the most promising one; hence this package has been used in this work. We can observe the performances of the public Sentiment Analyzer (SA) packages in the Tables I, II and III.

IV. PERFORMANCE ANALYSIS OF PROPOSED APPROACH

For the performance analysis of the proposed approach, ROUGE evaluation measure has been used. Chin-Yew Lin [13] proposed ROUGE measure that stands for Recall-Oriented Understudy for Gisting Evaluation. A public implementation of this approach has been used that is available in PyPI [3].

A. Performance of Approach 01

The performance of *Approach 01* of our proposed summarization technique is shown in Table IV.

TABLE IV
PERFORMANCE OF *Approach 01*

Article Category	ROUGE 1 Precision	ROUGE 1 Recall
Business	0.833	0.736
Entertainment	0.777	0.706
Politics	0.804	0.746
Sport	0.764	0.691
Tech	0.821	0.758
Average	0.799	0.727

B. Performance of Approach 02

The performance of *Approach 02* of our proposed summarization technique is shown in Table V.

TABLE V
PERFORMANCE OF *Approach 02*

Article Category	ROUGE 1 Precision	ROUGE 1 Recall
Business	0.806	0.682
Entertainment	0.74	0.639
Politics	0.792	0.689
Sport	0.76	0.659
Tech	0.773	0.658
Average	0.774	0.665

C. Analysis of Performance

We have observed that *Approach 01* has performed better than *Approach 02*. Hence, we can say that while summarizing, focus should be given to the sentences whose sentiment polarity is same as the overall document. We can also state that the proposed hybrid approach works very to generate automatic summary from a newspaper article.

V. CONCLUSION AND FUTURE WORKS

Since online newspaper platforms have become very popular, summarization of articles has become a very important task. This can help to bring out important information in a very short time. The proposed approach is promising and has achieved good performance scores through evaluation. In future, we intend to work with more ranking approaches and experiment with more datasets of newspaper articles not only of English language but also of many other languages.

REFERENCES

[1] Reda Elbarougy, Gamal Behery, and Akram El Khatib: Extractive arabic text summarization using modified PageRank algorithm. Egyptian Informatics Journal 21(2), 73-81 (2020)

[2] K. Usha Manjari: Extractive Summarization of Telugu Documents using TextRank Algorithm. In: 2020 Fourth International Conference on I-SMAC, Palladam, India (2020)

[3] Python Package Index (PyPI), https://pypi.org/. Last Accessed 12 September 2022

[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 (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 (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, Noida, India (2020)

[7] Chih-Fong Tsai, Kuanchin Chen, Ya-Han Hu, and Wei-Kai Chen, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," *Tourism Management*, vol 80, 2020.

[8] S. M. Meena, M. P. Ramkumar, R. E. Asmitha, and Emil Selvan G SR, "Text Summarization Using Text Frequency Ranking Sentence Prediction," in *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, Chennai, India, 28-29 September 2020.

[9] Sarunya Nathonghor, and Duangdao Wichadakul, "Extractive Text Summarization for Thai Travel News Based on Keyword Scored in Thai Language," in *Proceedings of the 2020 2nd International Conference on Information Technology and Computer Communications*, Kuala Lumpur Malaysia, August 2020.

[10] Reddy Naidu, Santosh Kumar Bharti, Korra Sathya Babu, and Ramesh Kumar Mohapatra, "Text summarization with automatic keyword extraction in telugu e-newspapers," *Smart computing and informatics*, pp. 555-564, 2018.

[11] Kaggle, https://www.kaggle.com/datasets. Last Accessed 10 September 2022

[12] S. Rose, D. Engel, N. Cramer, and W. Cowley: Automatic Keyword Extraction from Individual Documents. Text Mining: Applications and Theory, 1-20, (2010)

[13] Chin-Yew Lin: ROUGE: A Package for Automatic Evaluation of summaries. In:Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain (2004)

[14] GloVe, https://nlp.stanford.edu/projects/glove/. Last Accessed 10 September 2022