e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

TEXT SUMMARIZATION OF NEWS ARTICLE USING EXTRACTIVE TECHNIQUE

Suchit Balaso Kharade[*1], Vivek Ashok Kale[*2], Ketan Yashwant Lonkar[*3], Dr. R.N. Atole[*4]

[*1,2,3,4]Shivnagar Vidya Prasark Mandal College Of Engineering Malegaon(BK), India.

ABSTRACT

The Summarization of News Articles using Extractive Techniques involves a systematic approach to distill essential information from a given news article. This text summarization approach for news articles involves a multi-step process. Initially, preprocessing techniques are applied to clean and structure the text, encompassing tasks such as removing HTML tags, converting text to lowercase, tokenization, and stopword removal. A Word2Vec model is then trained on the preprocessed text, enabling the conversion of words into semantically meaningful embeddings. The core of the summarization process employs the TextRank algorithm, utilizing sentence embeddings derived from the Word2Vec model to construct a similarity matrix and represent sentences in a graph. The ranking produced by TextRank determines the most important sentences, forming the extractive summary. Postprocessing steps, including filtering based on sentence length and redundancy removal, enhance the coherence and readability of the final summary. This comprehensive approach facilitates the extraction of key information from news articles, providing concise and informative summaries.

I. INTRODUCTION

In the age of information overload, the ability to distill vast amounts of textual data into concise and informative summaries is paramount. This paper introduces a robust approach to text summarization for news articles, leveraging a multi-faceted methodology encompassing preprocessing, Word2Vec modeling, extractive summarization using TextRank, and postprocessing techniques.

The initial stage involves meticulous preprocessing, where the raw text undergoes cleaning procedures to remove extraneous elements like HTML tags and special characters. Tokenization and stopword removal further refine the text, preparing it for subsequent analysis. The incorporation of a Word2Vec model proves crucial in this process, as it captures the semantic relationships between words, providing a rich embedding representation for each term in the news article.

The heart of the summarization process lies in the application of the TextRank algorithm, a graph-based approach that relies on the cosine similarity between sentence embeddings derived from the Word2Vec model. The resulting similarity matrix constructs a representation of sentences as nodes and their relationships as edges, facilitating the identification of key sentences based on their importance in the network. These highly ranked sentences form the extractive summary, effectively distilling the essence of the news article.

To enhance the coherence and readability of the summary, postprocessing steps are implemented. Sentence length filtering eliminates overly short or long sentences, contributing to the overall flow of the summary. Redundancy removal ensures that the extractive summary is devoid of repetitive information, providing a concise and focused representation of the original news article.

This comprehensive approach not only addresses the challenges of information overload but also empowers users to quickly grasp the core content of news articles, making it a valuable tool for efficient information consumption in today's fast-paced world.

II. LITERATURE SURVEY

1) Title: Graph-based abstractive biomedical text summarization Publication:
Journal of Biomedical Informatics
Summary: Summarization is the process of compressing a text to obtain its important informative parts. In recent years, various methods have been presented to extract important parts of textual documents to present them in a summarized form. The first challenge of these methods is to detect the concepts that well convey the main topic of the text and extract sentences that better describe these essential concepts. The second challenge is the correct interpretation of the essential concepts to generate new paraphrased sentences such that they are not exactly the same as the sentences in the main text. The first challenge has been addressed by many

researchers. However, the second one is still in progress. In this study, we focus on the abstractive summarization of biomedical documents. In this regard, for the first challenge, a new method is presented based on the graph generation and frequent itemset mining for generating extractive summaries by considering the concepts within the biomedical documents. Then, to address the second challenge, a transfer learning-based method is used to generate abstractive summarizations from extractive summaries. The efficiency of the proposed solution has been evaluated by conducting several experiments over BioMed Central and NLM's PubMed datasets.

2) Title: Briefing of Textual Information using Text Rank

Publication: Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023) IEEE Xplore Part Number: CFP23AZ5-ART; ISBN: 978-1-6654-9199-0

Summary: The world is very much advanced with the abundant growth of technology and the way we communicate is rapidly changing with it. Lot of information is being generated day by day. It is important to extract the useful information from it. Text Summarization is one of the solutions to it. Different statistical methods like TFIDF, TF are used for extracting the summary. These statistical methods focus mainly on most frequently occurred words and less on importance of sentences. There are many limitations of TF-IDF method like it is based on bag of words model, so it does not capture position in text, semantics, co-occurrences in different documents. The other limitation is it assigns low values to words that are relatively important. So, we proposed a system that uses a graph-based approach which extracts the significant sentences from the given text. The proposed system provides extractive summary, abstractive summary and keywords for single document. The keywords extracted from the summary helps in understanding the main idea of the document. The proposed model helps users in providing summary with important points and saves user time.

3) Title: A Comparative Study of Opinion Summarization Techniques

Publication: IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS Author:Surbhi Bhatia

Summary: In the Web 3.0 platforms, enormous amount of information is shared whereby individuals express their thoughts and opinions and learn from others' experiences.

Many e-commerce websites provide service of posting opinionated reviews to allow consumers post their opinions using free text. Examples of these e-commerce websites include eBay, Amazon, and Yahoo shopping. Summarizing text is taken as an interesting task of Natural Language Processing (NLP). The proposed work presents a comparative study of different techniques used for opinion summarization. Extractive approach uses the principle of principal component analysis (PCA). The work includes the application of PCA in summarization of text by reducing the number of dimensions in data (aspects) and relatively finding the summary of the reviews on ranking the most relevant ones, according to the prime aspects without any loss of information respective of a particular domain. The analysis is conducted on the standard Opinosis data set and comparison is made between both of the techniques to discuss which method generates more coherent and complete summary.

III. RELATED WORK

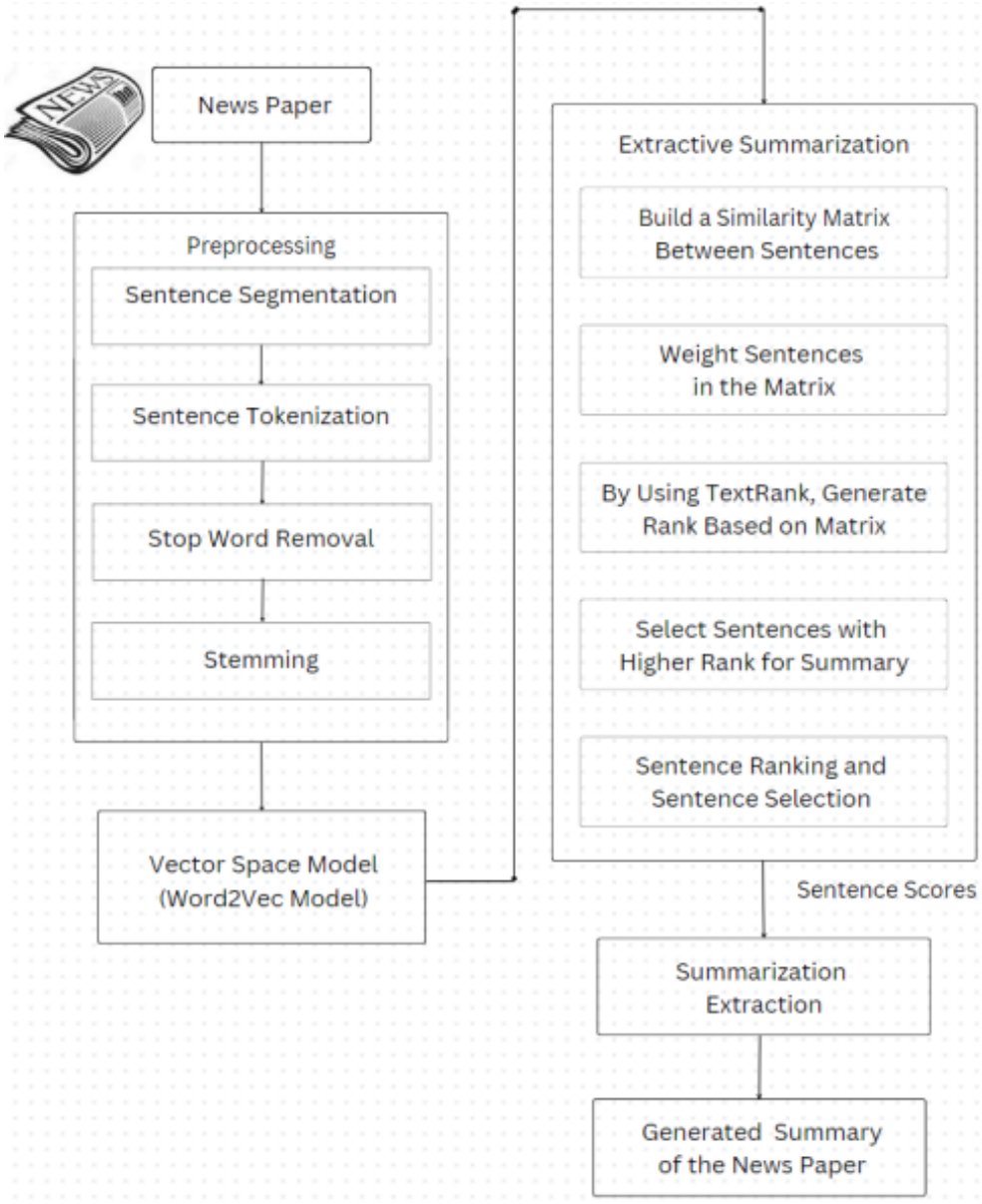
Summarizing news articles using extractive techniques has been a popular area of research in the field of natural language processing. Extractive summarization involves selecting and extracting sentences or phrases from the source text to create a concise summary. TextRank Algorithm: TextRank is an unsupervised extractive summarization technique inspired by PageRank. It treats sentences as nodes in a graph and uses the relationships between them to score and select the most important sentences for the summary.

IV. OBJECTIVE

The primary objective of this text summarization approach is to distill essential information from news articles efficiently. Through meticulous preprocessing, the raw text is refined for analysis. The integration of a Word2Vec model aids in capturing semantic relationships, enhancing the representation of words. The utilization of TextRank for extractive summarization enables the identification of key sentences, forming a concise summary. Postprocessing steps further refine the summary for coherence and readability, contributing to an effective and informative news article summarization process.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

V. SYSTEM ARCHITECTURE



VI. WORKING

1. Preprocessing:

a. Text Cleaning:

- Remove special characters, HTML tags and unnecessary symbols.
- Convert text to lowercase.

b. Tokenization:

- Break the text into individual words or phrases (tokens).

c. Stopword Removal:

- Eliminate common words that do not contribute much to the overall meaning.

d. Lemmatization/Stemming:

- Reduce words to their base or root form to capture their core meaning.

2. Word2Vec Model:

a. Training the Model:

- Train a Word2Vec model on your preprocessed text data.
- Word2Vec embeddings captures the semantic relationships between the words

b. Embedding Representation:

- Represent each word in the text with its corresponding Word2Vec embedding.

3. Extractive Summarization (TextRank):

a. Sentence Embeddings:

- Compute sentence embeddings by averaging or combining the Word2Vec embeddings of the words in each sentence.

b. Similarity Matrix:

- Create a similarity matrix based on the cosine similarity between sentence embeddings.

c. Graph-based Representation:

- Represent sentences as nodes and their similarities as edges in a graph.

d. TextRank Algorithm:

- Apply the TextRank algorithm to rank sentences based on their importance in the graph.

e. Extracting Summary:

- Extract the top-ranked sentences as the summary.

4. Postprocessing:

a. Sentence Length Filtering:

- Remove extremely short or long sentences to improve the coherence of the summary.

b. Redundancy Removal:

- Identify and remove redundant information to avoid repetition.

c. Summary Refinement:

- Fine-tune the summary for fluency, coherence, and readability.

VII. FUTURE SCOPE

The scope of Summarization of News Articles using Extractive Techniques is broad and encompasses various domains, offering numerous opportunities and addressing challenges in the evolving landscape of information consumption. the scope of Summarization of News Articles using Extractive Techniques is dynamic, spanning various fields and continuously evolving as technology advances. Its applications range from personal information management to broader societal impacts, contributing to a more efficient and accessible news consumption experience.

VIII. CONCLUSION

In conclusion, the proposed text summarization approach, integrating preprocessing, Word2Vec modeling, TextRank-based extractive summarization, and postprocessing techniques, emerges as a potent solution for distilling key insights from news articles. Through meticulous preprocessing, the raw text is refined for analysis, while the Word2Vec model captures semantic nuances, providing meaningful embeddings for words. The TextRank algorithm, leveraging these embeddings, effectively identifies and ranks important sentences, forming a coherent and informative extractive summary.

Postprocessing steps, including sentence length filtering and redundancy removal, contribute to the refinement of the summary, ensuring clarity and conciseness. This comprehensive methodology not only addresses the challenges of information overload but also enhances the efficiency of information consumption, allowing users to quickly comprehend the essence of news articles. The synergy of these techniques offers a robust solution for text summarization, contributing to the facilitation of rapid and insightful content assimilation in the dynamic landscape of news consumption.

IX. REFERENCES

[1] Z. Jalil, J. A. Nasir, and M. Nasir, Extractive MultiDocument Summarization: A Review of Progress in the Last Decade, IEEE Access, vol. 9, pp. 130928-130946, 2021, DOI: 10.1109/ACCESS.2021.3112496.

[2] R. Mishra, J. Bian, M. Fiszman, C.R. Weir, S. Jonnalagadda, J. Mostafa, G. Del Fiol, Text summarization in the biomedical domain: a systematic review of recent research, J. Biomed. Inform. 52 (2014) 457–467.

[3] M.N. Azadani, N. Ghadiri, E. Davoodijam, Graph-based biomedical text summarization: an itemset mining and sentence clustering approach, J. Biomed. Inform. 84 (2018) 42–58.

[4] J.C. Cheung, Comparing abstractive and extractive summarization of evaluative text: controversiality and content selection, B. Sc.(Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia, vol. 47, 2008.

[5] Á. Hernández-Castañeda, R. A. García-Hernández, Y. Ledeneva and C. E. Millán-Hernández, Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords, IEEE Access, vol. 8, pp. 49896-49907, 2020, doi: 10.1109/ACCESS.2020.2980226.

[6] Vinit Aghama, Dr.V.K.Shandilyab. "A Survey Paper on Extractive and Abstractive Techniques in Automatic Text Summarization", International Journal of Research Publication and Reviews Vol (2) Issue (4) (2021) Page 619-625..