News Text Summarization Using Latent Semantic Analysis (LSA) Algorithm

Manisha M. Langote[1] Dr. Ranjit Gawande[2]
[1,2]Department of Computer Engineering
[1,2]Matoshri College of Engineering & Research Center, Nashik, SPPU Pune University, Pune, India

Abstract — In proposed work, we successfully implemented a news text summarization system using Natural Language Processing (NLP) techniques and the Latent Semantic Analysis (LSA) algorithm. The purpose of our project was to extract important information from a large volume of news articles and present it in a concise and easily understandable manner. To achieve this, we utilized the LSA algorithm, which is known for its ability to capture the underlying semantic structure of text. LSA employs a mathematical model to analyse relationships between words in a document, creating a semantic representation where words with similar contexts are grouped together in a vector space. The LSA-based summarization process involved several steps. First, we pre-processed the news articles by removing stop words, punctuation, and other non-relevant elements. Then, we constructed a term-document matrix, where rows represented words and columns represented documents, with matrix values representing word frequencies. Next, we applied Singular Value Decomposition (SVD) to the term-document matrix. SVD helped reduce the matrix's dimensionality by identifying the most important latent semantic concepts. This resulted in a lower-dimensional representation that captured the essential information. Finally, we identified the most important sentences in the news articles by measuring the cosine similarity between each sentence and the summary. Sentences with the highest cosine similarity scores were selected as summary sentences. The proposed system demonstrated the effectiveness of the LSA algorithm for news text summarization. By capturing the semantic structure of the text, it generated summaries that allowed users to understand the key points of a news article quickly and easily. Our implementation had practical applications for content recommendation systems, news aggregation platforms, and personalized news feeds. However, it is important to acknowledge the limitations of the LSA algorithm. It may struggle with handling idiomatic expressions and can be sensitive to the quality of the input data. These considerations highlight the need for ongoing research and development to enhance the performance and robustness of news text summarization systems.

Keywords: News Text Summarization, News Aggregation Platforms, Latent Semantic Analysis (LSA), Semantic Structure, Singular Value Decomposition (SVD), Preprocessed, Stop Words

I. INTRODUCTION

Text summarization is the process of automatically condensing a large body of text into a shorter version while retaining its most important information. It is a crucial task in natural language processing and is used in various applications, including news article summarization, document summarization, and email summarization. There are various techniques for text summarization, including extraction-based summarization, abstraction-based summarization, and latent semantic analysis (LSA)-based summarization. LSA is a mathematical technique that is commonly used for information retrieval and text summarization. It is a type of unsupervised learning that extracts latent information from a corpus of text documents by analyzing their co-occurrence patterns. The LSA algorithm works by constructing a matrix of term-document frequencies, which represents the occurrence of each word in each document. This matrix is then transformed into a lower-dimensional space using singular value decomposition (SVD), which identifies the most important latent features of the data. These latent features correspond to the underlying concepts in the text, such as topics or themes.

In the context of text summarization, LSA can be used to identify the most important sentences in a document by representing the document as a vector of latent features and then ranking the sentences based on their similarity to this vector. The top-ranked sentences can then be selected to form a summary of the document. LSA-based summarization has been shown to produce summaries that are more informative and coherent than those produced by other methods, such as extraction-based summarization. One of the advantages of LSA-based summarization is that it can capture the underlying semantics of the text, allowing it to identify important concepts and relationships that may not be apparent from a simple word frequency analysis. Additionally, LSA-based summarization can handle documents with varying lengths and structures, making it suitable for a wide range of applications. However, LSA-based summarization also has some limitations. One of the main challenges is determining the optimal number of latent features to use, as this can have a significant impact on the quality of the summary. Another challenge is dealing with noisy or irrelevant information in the text, which can negatively impact the summarization performance.

Despite these challenges, LSA-based summarization remains a popular and effective technique for text summarization. Researchers continue to work on improving the algorithm and addressing its limitations, with the aim of developing more accurate and efficient summarization methods.

II. RELATED WORK

This research [1] work comprises an automatic text categorization and summarization approach to analyze the structure of input text. In this work, a text analyzer is developed to derive the structure of the input text using the rule reduction technique in three stages namely, Token Creation, Feature Identification and Categorization, and Summarization. This analyzer is tested with sample input texts and gives noteworthy results. Extensive experimentation validates the selection of parameters and the

efficacy of our approach for text classification. This work can be expanded and used in many practical applications, including indexing for document retrieval, organizing, and maintaining large catalogs of Web resources, automatically extracting metadata, and Word sense disambiguation, etc.

In this paper [2], a novel statistical method to perform an extractive text summarization on a single document is demonstrated. The method extraction of sentences, which gives the idea of the input text in a short form, is presented. Sentences are ranked by assigning weights and they are ranked based on their weights. Highly ranked sentences are extracted from the input document, so it extracts important sentences that direct to a high-quality summary of the input document and stores the summary as audio.

This paper [3] is a survey on the various types of text summarization techniques starting from the basic to the advanced techniques. According to this survey, the seq2seq model along with the LSTM and attention mechanism is used for increased accuracy.

This review paper [4] presents various approaches to generate a summary of huge texts. Various papers have been studied for different methods that have been used so far for text summarization. Mostly, the methods described in this paper produce Abstractive (ABS) or Extractive (EXT) summaries of text documents. Query-based summarization techniques are also discussed. The paper mostly discusses the structured-based and semantic-based approaches for the summarization of text documents. Various datasets were used to test the summaries produced by these models, such as the CNN corpus, DUC2000, single and multiple text documents, etc. We have studied these methods and the tendencies, achievements, past work, and future scope of them in text summarization as well as other fields.

III. Proposed System

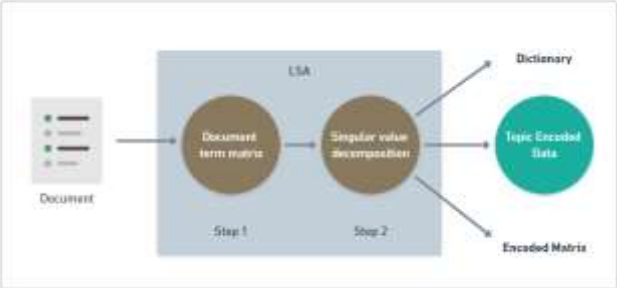

Fig. 1: Architecture of Proposed System

The proposed solution will help in many ways like reducing the reading time of long documents while buying products online, and summaries of product reviews making the selection process easier. Extractive summarization is basically creating a summary based on strictly what you get in the text. It can be compared to copying down the main points of a text without any modification to those points and rearranging the order of those points and the grammar to make more sense out of the summary.

The algorithm for LSA consists of three major steps:

– Input matrix creation: The input document is represented as a matrix to understand and perform calculations on it. Thus, a document term matrix is generated. Cells are used to represent the importance of words in sentences. Different approaches can be used for filling out the cell values. There are different approaches to filling out the cell values.
– Singular Value decomposition (SVD): In this step, we perform the singular value decomposition on the generated document term matrix. SVD is an algebraic method that can model relationships among words/phrases and sentences. The basic idea behind SVD is that the document term matrix can be represented as points in Euclidean space known as vectors. These vectors are used to display the documents or sentences in our case in this space. Besides having the capability of modeling relationships among words and sentences, SVD has the capability of noise reduction, which helps to improve accuracy.
– Sentence Selection: Using the results of SVD different algorithms are used to select important sentences. Here we have used the Topic method to extract concepts and sub-concepts from the SVD calculations which are called topics of the input document. These topics can be sub-topics, and then the sentences are collected from the main topics.

The high number of common words among sentences indicates that the sentences are semantically related. The meaning of a sentence is decided using the word it contains and the meaning of words is decided using the sentences that contain the word. Dictionary and Encoding matrix are the by-products obtained during the execution of LSA processing. A dictionary is a set of all words that occur at least once in our document. While the encoded data represents words of our sentences in terms of their individual strengths. This strength helps us determine the exact effect of each word in our sentence/document.

IV. Result and Analysis

The LSA algorithm can generate summaries of varying lengths, depending on the desired level of detail. The length of the generated summaries can be evaluated to ensure they align with the expected summary length.

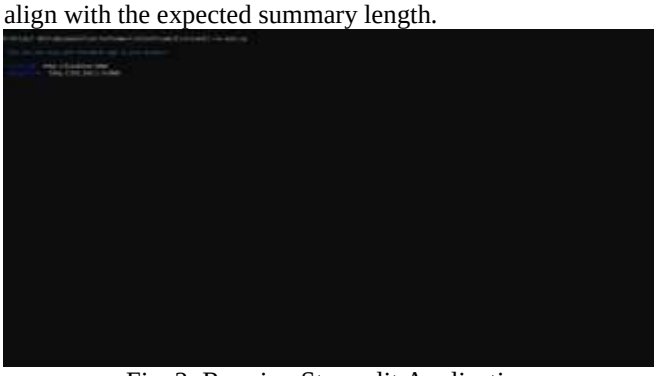

Fig. 2: Running Streamlit Application

The LSA algorithm can be evaluated based on its ability to rank the importance of sentences accurately. Higher-ranked sentences should contain more crucial information compared to lower-ranked sentences.

Fig. 3: Result of the System

User feedback and subjective evaluations can provide valuable insights into the perceived quality of the summaries. Collecting feedback from users, such as journalists or readers, can help identify areas of improvement and assess the overall usefulness of the LSA-based summarization approach.

V. Conclusion

The Proposed work focused on utilizing Natural Language Processing (NLP) and Machine Learning techniques with the Latent Semantic Analysis (LSA) algorithm for News Text Summarization. The LSA algorithm was employed to extract latent semantic meaning from news articles and generate concise summaries. The LSA algorithm proved effective in identifying important concepts and themes within news articles. By leveraging co-occurrence patterns of words and applying singular value decomposition, the algorithm successfully captured the latent semantic structure of the document collection. The generated summaries exhibited varying lengths based on the desired level of detail, providing flexibility to suit different summarization requirements. However, it is crucial to ensure that the summaries align with the expected length and level of information coverage. The project highlighted the importance of comparing the LSA-based summarization results with reference summaries or human-generated summaries to evaluate the performance of the algorithm.

VI. Future Scope

One of the future improvements may be to apply the topic-focused summarization framework to news articles or blogs and to extend the work in the machine learning approaches. Topic-focused summaries of news articles would be a lot more accurate and valuable to users. It would be more interesting to work on topic modeling and summarization in the domain of social media in future.

VII. References

[1] C. Lakshmi Devasena and M. Hemalatha, "Automatic Text categorization and summarization using rule reduction," IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM - 2012), 2012, pp. 594-598.

[2] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, DOI: 10.1109 / IconDSC. 2019.8817040.

[3] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355

[4] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approach for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538.

[5] P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021

[6] X. -y. Jiang, X. -Z. Fan, Z. -F. Wang and K. -L. Jia, "Improving the Performance of Text Categorization Using Automatic Summarization," *2009 International Conference on Computer Modeling and Simulation*, 2009

[7] K. D. Garg, V. Khullar and A. K. Agarwal, "Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), 2021

[8] M. Ji, R. Fu, T. Xing and F. Yin, "Research on Text Summarization Generation Based on LSTM and Attention Mechanism," 2021 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), 2021

[9] S. R. Rahimi, A. T. Mozhdehi and M. Abdolahi, "An overview on extractive text summarization," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017

[10] E. Reategui, M. Klemann and M. D. Finco, "Using a Text Mining Tool to Support Text Summarization," 2012 IEEE 12th International Conference on Advanced Learning Technologies, 2012

[11] C. HARK, T. UÇKAN, E. SEYYARER and A. KARCI, "Graph-Based Suggestion for Text Summarization," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018

[12] C. Kwatra and K. Gupta, "Extractive and Abstractive Summarization for Hindi Text using Hierarchical Clustering," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2021

[13] P. S. Suryadjaja and R. Mandala, "Improving the Performance of the Extractive Text Summarization by a Novel Topic Modeling and Sentence Embedding Technique using SBERT," 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2021

[14] Zhang, R., Li, L., & Ji, H. (2018). Extractive summarization using bert embeddings. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 3725-3731). IEEE.

[15] Liu, P., Qiu, X., & Huang, X. (2019). Fine-grained opinion mining with recurrent neural networks for news

summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(2), 434-444.
[16] Zhang, X., Zhao, T., & LeCun, Y. (2018). Joint multi-label classification and multi-task learning with convolutional networks. In Proceedings of the 35th International Conference on Machine Learning (Vol. 80, pp. 5665-5674). IEEE.
[17] Yao, Z., Yang, Z., & Li, C. (2018). News summarization via unsupervised deep learning. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1302-1309). IEEE.
[18] Nallapati, R., Zhou, B., Dos Santos, C., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (pp. 280-290).
[19] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2015). A latent semantic model with convolutional-pooling structure for information retrieval. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 69-78).
[20] Li, C., Shi, Y., Yan, Y., Qi, J., & Chen, H. (2019). An enhanced extractive summarization framework with document-level semantic relations. IEEE Access, 7, 29821-29830.
[21] Xiao, Y., Yao, Z., Yang, Z., Li, C., & Xu, J. (2020). Attention-based neural networks for extractive summarization. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1615-1628.
[22] Ma, L., Huang, J., & Xie, X. (2017). Neural network-based extractive text summarization with application to news. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4), 754-764.
[23] Nayeem, T., Roy, S., & Al-Maadeed, S. (2019). Multi-document extractive text summarization using graph-based approach. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 3465-3472). IEEE.
[24] Chakraborty, D., & Basak, J. (2020). A novel approach for news summarization using deep learning. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

