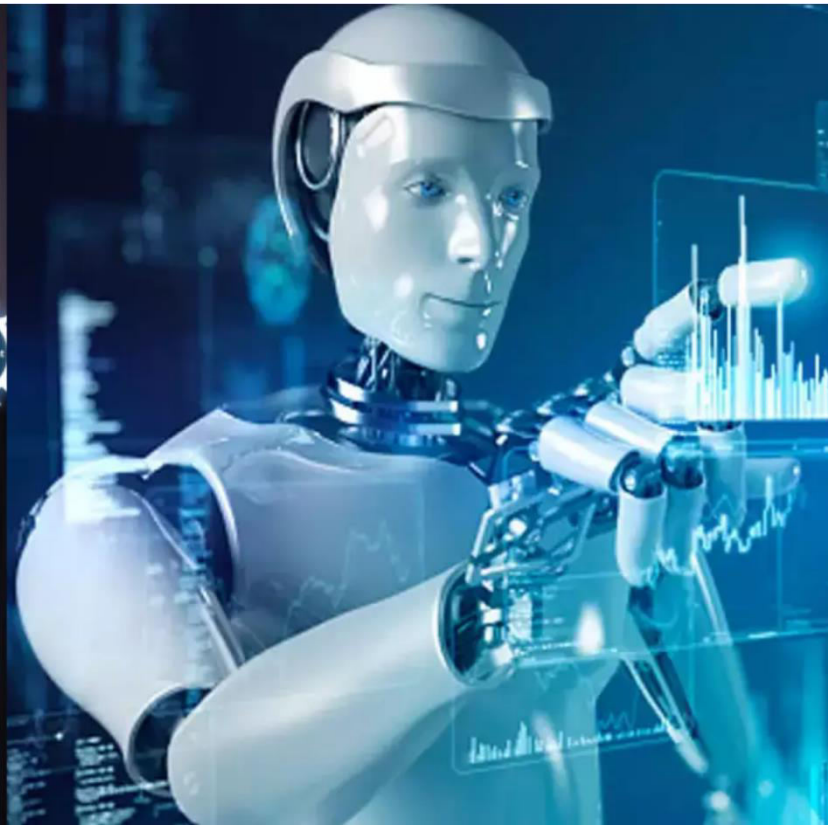
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404647|

Text Summarization System using Natural Language Processing

Pradip Uddhav Mache, Prof. Nanda Kulkarni

Department of Computer Engineering, Siddhant College of Engineering, Pune, India

ABSTRACT: Automatic text summarization, an essential natural language processing application, uses machine learning algorithms to compress a given textual content into a shorter model. As media material transmission over the Internet continues to grow at an exponential rate, text summarization using neural networks from asynchronous text combinations is becoming increasingly important. Using natural language processing (NLP) principles, this study presents a framework for investigating the complicated information included in multi-modal statistics as well as refining the current text summarization characteristics. The core premise is to fill semantic gaps between various sorts of content. The following stage use multi-modal topic modeling to provide a summary of pertinent information.Finally, all multi-modal factors are considered in order to give a textual summary that maximizes the information's relevance, non-redundancy, believability, and scope through the allocation of submodular features.

KEYWORDS: word vectors, word analogies, fast text, Integer linear programming, text summarising, natural language processing.

I. INTRODUCTION

Nowadays, there are numerous documents or information available for any given sector. There are numerous sources from which we might obtain a wealth of information relevant to our subject of search. Much information is available from different sources, including the internet. However, we all know that a vast amount of information cannot always be examined or used. So, an exact amount of information is always considered, and that information is extracted from the original document, which is massive in size. In other words, we extracted the summary of the main document.

A summary of any document is described as a collection of vital material compiled from brief assertions that account for the main points of the original document. As a result, text summarization is the process of separating or extracting useful information from a huge material. It is the process of shortening a written content using various technologies and approaches to provide a cohesive summary that includes the original document's main ideas. There are several approaches for carrying out the summary procedure. While most summarizing systems focus primarily on natural language processing (NLP), the ability to jointly optimize the quality of the summary with the use of automatic speech recognition (ASR) and computer vision.

(CV) processing systems are largely disregarded. In contrast, multimedia data associated with a news event (i.e., a news topic) are typically asynchronous in real life. As a result, comprehending the semantics of information presents a significant difficulty for text summarization. In this paper, we offer a system that may provide users with textual summaries to assist them get the gist of asynchronous data in a short amount of time without having to read entire documents. The goal of this research is to combine NLP and neural network approaches to develop a new framework for mining the rich information contained in multimodal data in order to improve the quality of text summarization.

Summarization is the process of condensing a lengthy piece of content into a shorter version that preserves the most important information. There are two sorts of summarization techniques: extractive and abstractive. Extractive procedures generate summaries primarily from parts (usually whole sentences) retrieved directly from the source material, whereas abstractive methods may generate new words and phrases not found in the source text, as a human-written abstract might. Because replicating large chunks of text from the original manuscript gives baseline levels of correctness, the extractive method is simpler. In previous papers using extractive techniques, text summarization was divided into two subtasks: sentence scoring and sentence selection. Sentence scoring is a well-researched technique for assigning an important score to each sentence. Extractive methods generate summaries by recreating portions of the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404647|

source content (often full sentences), whereas abstractive approaches may generate new words or phrases that do not appear in the original document. Extractive summarization, often known as a sentence ranking or binary classification problem (i.e., sentences that are highest ranked or projected to be True are chosen as summaries), has gotten a lot of attention in the past. Content selection in summarization is often accomplished by sentence (and, on rare occasions, phrase) extraction. Despite the fact that deep learning models are an important component of both extractive and abstractive summarization systems, it is unclear how they pick content using simply word and sentence embedding data as input. One of the most difficult NLP tasks is summary, which is described as the act of producing a shorter version of a piece of text while retaining crucial contextual information.

The availability of large-scale datasets, on the other hand, is important to the performance of these models. Furthermore, the length of the articles and the variety of genres may contribute to their complexity. Because news articles have unique properties, systems trained purely on news may not be sufficiently generalised. Recent advances in machine learning have resulted in considerable improvements to text summarization. Large labelled summarization corpora, such as the CNN/Daily Mail dataset, have enabled the training of deep learning models with many parameters. They included the recurrent neural network (RNN) and Arif Ur Rahman, the assistant editor who handled the manuscript's examination and approval for publishing. Convolution neural networks (CNN) are widely used for text summarization. RNN is utilized in extractive techniques to determine sentence importance while also selecting representative sentences.

II. LITERATURE SURVEY

Abigail See et al: We showed in this research that a hybrid pointer generator architecture with coverage reduces repetition and inaccuracy. Our model performed noticeably better than the abstract state-of-the-art result when we tested it on a new and challenging lengthy text dataset. Although our model has many abstract talents, it is still in the early stages of developing the ability to reach higher levels of abstraction.

Qingyu Zhou et al: In this paper, we solve this issue by introducing a novel neural network architecture for extractive document summarization that simultaneously learns to score and choose sentences. Our approach combines sentence grading and selection into a single phase, which is the most obvious distinction from previous approaches. Every time it chooses a sentence, it rates it according to the current extraction state and the partial output summary. The findings of the ROUGE study show that the proposed combination sentence scoring and selection strategy works significantly better than the current segregated method.

Xingxing Zhang et al: In this research, we proposed a latent variable extractive summarization technique that directly exploits human summaries using a phrase compression model. In experiments, the suggested method performs better than a strong extractive model, but using the compression model on our extractive system's output yields less satisfactory results. In the future, we plan to look into methods for training compression models especially for our summarizing task.

Chris Kedzie et al: Deep learning-based content selection algorithms for summarization were empirically investigated in this paper. Our results suggest that further work on sentence representation for summarization is needed, and that such models have substantial limitations in terms of their ability to learn strong features for this job.

Linqing Liu et al: In this paper, we proposed an adversarial method for abstractive text summarization. Our model was able to generate more abstract, readable, and varied summaries, according to experiments.

Jacob Devlin et al: Rich, unsupervised pre-training is a crucial component of many language comprehension systems, as demonstrated in this study by recent empirical advances in language models brought about by transfer learning. Specifically, these results demonstrate that even low-resource tasks can benefit from deep unidirectional topologies. We make a significant addition by applying these results to deep bidirectional architectures, which enable a single pre-trained model to handle a variety of NLP tasks.

T. Boongoen et al: Both traditional and more modern methods for cluster ensembles have been presented in this survey. The formal terminology used to define the problem comes first. The four primary categories of consensus clustering

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404647|

techniques are then thoroughly described with examples. The three main parts of a cluster ensemble framework—consensus function, representation and summarization, and ensemble generation—are then extended. Due to their enhanced capabilities, numerous cluster ensemble techniques have been used to a variety of applications and domain difficulties.

Mahnaz Koupaee et al: In this work, we introduce WikiHow, a brand-new extensive summary dataset composed of several articles from the WikiHow knowledge base. The research's descriptions of WikiHow's features could provide summarization algorithms with more difficulties. We anticipate that scholars will be interested in the new dataset as a promising alternative for assessing their systems.

Edouard Grave et al: This study includes a quick language identifier that can recognize 176 languages, word vectors trained on Wikipedia and the Common Crawl, and three additional analogy datasets to assess these models. We show how to generate high-quality word vectors by examining the effects of various hyperparameters on the performance of the trained models. We also demonstrate that using common crawl data can produce superior models for languages with low Wikipedia coverage and models with higher coverage, despite its noise. Lastly, we discover that compared to other languages, the quality of the generated word vectors for low-resource languages like Hindi is significantly lower. In the future, we would like to investigate more methods for enhancing the caliber of models for these languages.

Zhilin Yang et al: XLNet is a generalized AR pretraining technique used in this study that uses a permutation language modeling goal to combine the advantages of AR and AE techniques. With the careful design of the two-stream attention mechanism and the incorporation of Transformer-XL, XLNet's neural architecture was created to complement the AR objective. XLNet provides notable enhancements over prior pretraining goals on a range of tasks.

III. PROPOSED SYSTEM

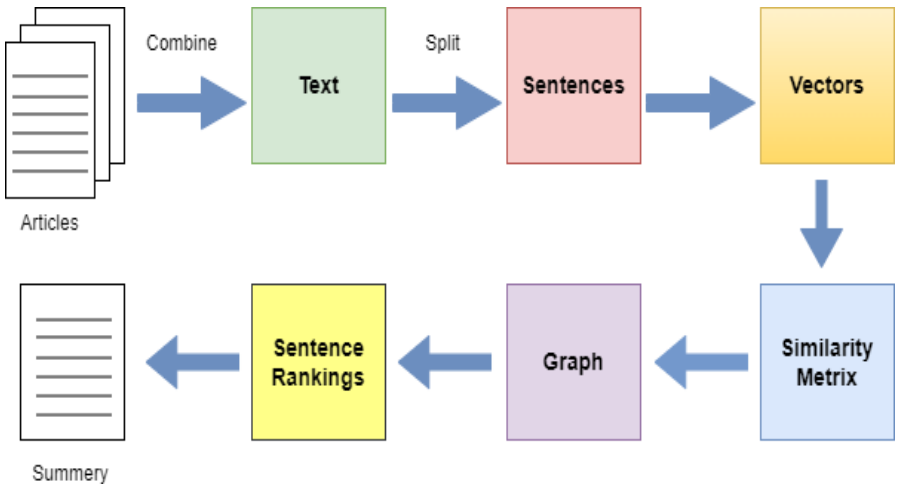


Fig. 1. Proposed System Architecture

The procedure consists of two steps: document representation at the beginning and sample sentence selection at the conclusion. A document is represented as a continuous vector using distributed vectors that have been pre-trained using phrase embedding models. Representative phrases for the summary are then selected after the sentence importance score is evaluated using ILP and PCA. To further improve model performance, a variety of pre-trained sentence representations were employed in an ensemble. Concatenating all of the text in the articles would be the first step. Next, divide the text into separate sentences. The next stage will be to determine each sentence's vector representation. After that, sentence vector similarity is computed and recorded in a matrix.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404647|

Algorithm

ERT - Bidirectional Encoder Representations from Transformers

ERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that can understand the context of words in a sentence by looking both before and after a given word. This bidirectional approach allows BERT to grasp the nuanced meaning of words depending on their context.

SciBERT is essentially a version of BERT tailored for the scientific domain, equipped with a deep understanding of scientific language and context. By using a custom vocabulary and training on scientific literature, SciBERT serves as a foundation for tasks that involve understanding, analyzing, and summarizing research articles. Its integration with transformer-based models like Graph Transformer Networks (GTN) allows it to contribute significantly to complex tasks like generating abstractive summaries that capture the core ideas and relationships within scientific documents.

IV. RESULTS AND DISCUSSION

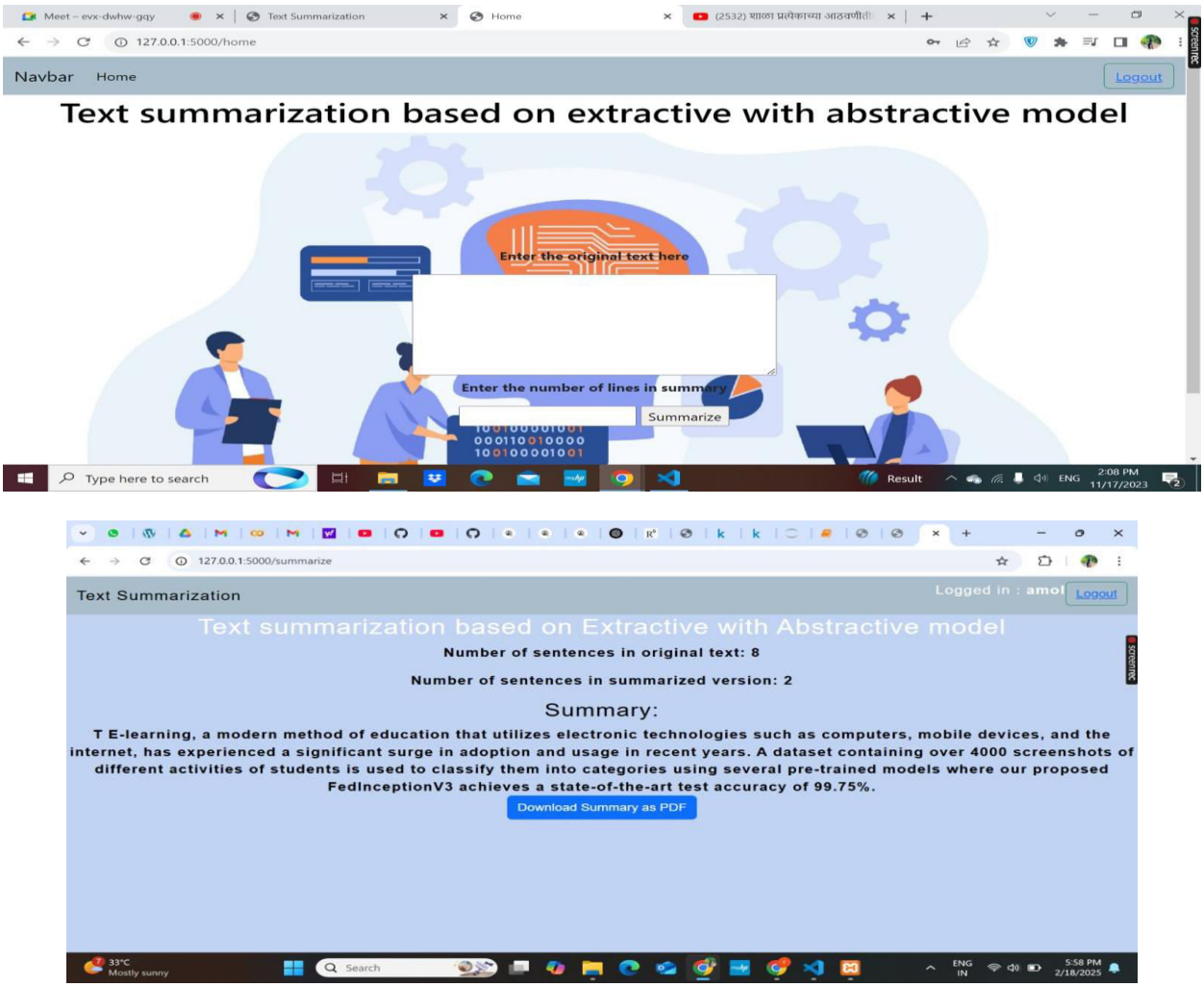


Figure 2. Results Screenshots

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404647|

V. CONCLUSION

In this research, we offer a model for summarization that does not require training of parameters. To build a deep representation of documents, we employed positional encoding, self-attention, and pre-trained sentence vectors. In order to choose an appropriate number of summary sentences and determine the relevance score of each source sentence, we next extracted PS vectors from a document that retain as much information as feasible. In order to generate the final summarization results, we built an ensemble model based on different models with different pre-trained sentence embedding vectors. We verified that the performance of the proposed model (without training examples) was on par with a cutting-edge supervised model based on machine learning.

VI. FUTURE SCOPE

In future research we will focus on multi-document and cross-document text summarization.

REFERENCES

[1] MYEONGJUN JANG AND PILSUNG KANG,"Learning-Free Unsupervised Extractive Summeriza- tion Model ," 2021, arXiv:9321308 [Online].Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=9321308
[2] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Sum- marization with pointer-generator networks," 2017, arXiv:1704.04368. [Online].Available: http://arxiv.org/abs/1704.04368.
[3] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "Neural document summarization by jointly learning to score and select sen- tences," 2018, arXiv:1807.02305. [Online]. Available: http://arxiv.org/ abs/1807.02305.
[4] X. Zhang, M. Lapata, F. Wei, and M. Zhou, "Neural latent extractive document summarization," 2018, arXiv:1808.07187. [Online]. Avail-able:http://arxiv.org/abs/1808.07187.
[5] C. Kedzie, K. McKeown, and H. Daume, "Content selection in deep learning models of summarization," 2018, arXiv:1810.12343. [Online].
[6] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, "Generative adversarial network for abstractive text summarization," in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–2.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre training of deep bidirectional transformers for language understanding," 2018,arXiv:1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.
[8] T. Boongoen and N. Iam-On, "Cluster ensembles: A survey of ap- proaches with recent extensions and applications," Comput. Sci. Rev., vol. 28, pp. 1–25, May 2018.
[9] M. Koupaee and W. Y. Wang, "WikiHow: A large scale text summarization dataset," 2018, arXiv:1810.09305. [Online]. Available: http://arxiv.org/abs/1810.09305.
[10] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in Proc. Int. Conf. Lang. Resour. Eval. (LREC), 2018, pp. 1–3.
[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and
[12] Q. V. Le,"XlNet: Generalized autoregressive pretraining for language understanding," in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 5753–5763.
[13] C. Kedzie, K. McKeown, and H. Daume, "Content selection in deep learning models of summarization," 2018, arXiv:1810.12343. [Online].
[14] Available: http://arxiv.org/abs/1810.12343
[15] Z. Li, Z. Peng, S. Tang, C. Zhang, and H. Ma, "Text summarization method based on double attention pointer network," IEEE Access, vol. 8, pp. 11279–11288, 2020
[16] J. Tan, X. Wan, and J. Xiao, "Abstractive document summarization with a graph-based attentional neural model," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, Jul. 2017, pp. 1171–1181.
[17] H. Kim and S. Lee, "A context based coverage model for abstractive document summarization," in Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC), Oct. 2019, pp. 1129–1132.
[18] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, "Generative adversarial network for abstractive text summarization," in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–2.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details