COMPUTERIZED TEXT SUMMARIZATION

Vijay Sonawane[1], Rishi Chaudhary [2], Sourav Savant [3], Pawan Bhosle [4], Akash Singh [5],

[1]Assistant Professor, [2,3,4,5]Department of Computer Engineering, JSPM's Bhivarabai Sawant Institute of Technology Research, Wagholi, Pune.

Abstract: One essential aspect of natural language processing (NLP) is automated text summarization activity that is essential for managing the massive amount of textual information that is currently available in the digital age. The fundamental elements of automated text summarization, including its importance, techniques, difficulties, and applications, are briefly covered in this abstract. A need for effective ways to condense and convey textual information succinctly and efficiently has arisen as a result of the exponential growth of digital content. Techniques for text summarizing are designed to create automatically shorter copies of lengthy materials while preserving their main ideas. Information retrieval, news aggregation, document classification, and content creation are just a few of the uses for this technique. The extractive and abstractive subcategories of text summarization are examined in this abstract. Selecting is required for extractive summarization.

1. Introduction

Effective information retrieval and knowledge management are more important than ever in the information-rich world of today, when an enormous volume of text-based content is produced every day. Tools for computerized text summarization have become essential answers to this problem. These programs use Natural language processing (NLP) and artificial intelligence (AI) to automatically condense long texts, articles, or textual data into succinct summaries. People and organizations are now surrounded by more information than they can possibly consume as a result of the continual upsurge in the development of digital content. In this situation, computerized text summary plays a crucial role in helping readers sift through a sea of information to find the most insightful ideas, salient points, and crucial details. A trustworthy text summarizing tool can be your savior whether you're a student attempting to comprehend a long research paper, a professional keeping up with industry trends, or a researcher sifting through large quantities of literature.

The purpose of this introduction is to acquaint you with the idea of automated text summary by illuminating its significance, practicality, and the numerous ways it may make engagement with text content easier. These technologies can quickly process large quantities of text using sophisticated algorithms, making the important information easier to access and comprehend. At its core, a computerized text summary tool extracts the main ideas from a paper, web page, or other textual input and condenses them into a more manageable, coherent version. Text summarizing methods is split into two groups: abstractive and extractive. Extractive summarizing includes picking out and putting together words or phrases to produce a summary from the source content that is essentially verbatim. As opposed to this, abstractive summarization makes use of cutting-edge NLP approaches to produce summaries that are more creative and human-like, frequently reworking the text in a way that keeps the essential ideas while utilizing different terminology.

2. Need For the Project

Information Overload: We are inundated with a massive amount of text-based content every day in the digital age, including emails, social media posts, research papers, and news articles. Text summarization tools make it easier for people and organizations to sift through this enormous sea of data and extract the most crucial and pertinent information.

Time Efficiency: Long documents can take a while to read and understand. Users of text summarization tools can quickly understand the main ideas and important takeaways from a text without having to read it in its entirety. Professionals, researchers, and students who need to effectively review a lot of content will find this to be especially helpful.

Content Aggregation: For content aggregation and curation, text summarization is crucial. It is simpler for readers to scan headlines and summaries before reading the full articles when news aggregators, for instance, use summarization to produce concise versions of news articles.

Information Retrieval: Text summarization helps search engines and databases give users excerpts or summaries of web pages or documents in search results. This enables users to assess the content's suitability for their query quickly.

Language Translation: Summarization is an important step in the machine translation process because it allows for the condensing of lengthy sentences or paragraphs while maintaining the meaning that is most important in the target language.

Educational Aids: By offering condensed summaries of textbooks and academic papers, text summarization tools in educational contexts can aid students in understanding difficult concepts.

3. Problem Statement

"Develop a computerized text summarizing system that successfully deals with issues including information overload, diverse material, coherence, and usability. It should be flexible enough to accommodate many languages and information kinds, take ethical issues into account, and compete favorably with currently used summary techniques. In the era of plentiful information, the successful completion of this project will help to improve information retrieval effectiveness and user comprehension."

4. Existing System

Methodology

Data Collection and Preprocessing: Collect the documents or textual information that needs to be distilled. Tokenize the data to separate the text into sentences and phrases and remove noise such as special characters and unnecessary formatting.

Content Understanding: Examine the material to comprehend its context, major ideas, and organizational structure. This may entail determining crucial terms, entities, and the connections between them.

Sentence Scoring (Extractive Summarization): In extractive summarization, sentences are scored based on their relevance to the main content of the document. Common scoring methods used are: Term Frequency-Inverse Document Frequency (TF-IDF) scores. Graph-based algorithms like Text Rank or PageRank that consider sentence similarity and importance.

Sentence Ranking and Selection (Extractive Summarization): Sort sentences according to how well they performed, then choose the top-scoring phrases to use as the summary. The number of sentences that are chosen may be predetermined or determined by the user.

Abstraction (Abstractive Summarization): Abstractive summarization involves rewriting or rephrasing the content to create summaries that are more akin to what a human would say. In models of sequence-to-sequence, the decoder creates the summary whereas the encoder normally encodes the input text.

Evaluation: Utilize evaluation metrics like ROUGE that is Recall-Oriented Understudy Gisting Evaluation or BLEU which is Bilingual Evaluation Understudy to rate the effectiveness of the generated summary. These metrics gauge how closely the reference summary and the summary that was generated overlap and are similar to one another.

Tuning and Optimization: To enhance performance, fluency, and coherence, Use a training set to fine-tune the summarization model. dataset and optimization strategies.

Deployment: Any application or platform that requires summarization, such as a web service, chatbot, document management system, or other situation, should incorporate the summarization tool.

Maintenance and Updates: Maintain and update the summarization system frequently to reflect changes in the language, data sources, and user needs. This might entail updating the algorithms and retraining the model.

5. Suggested system

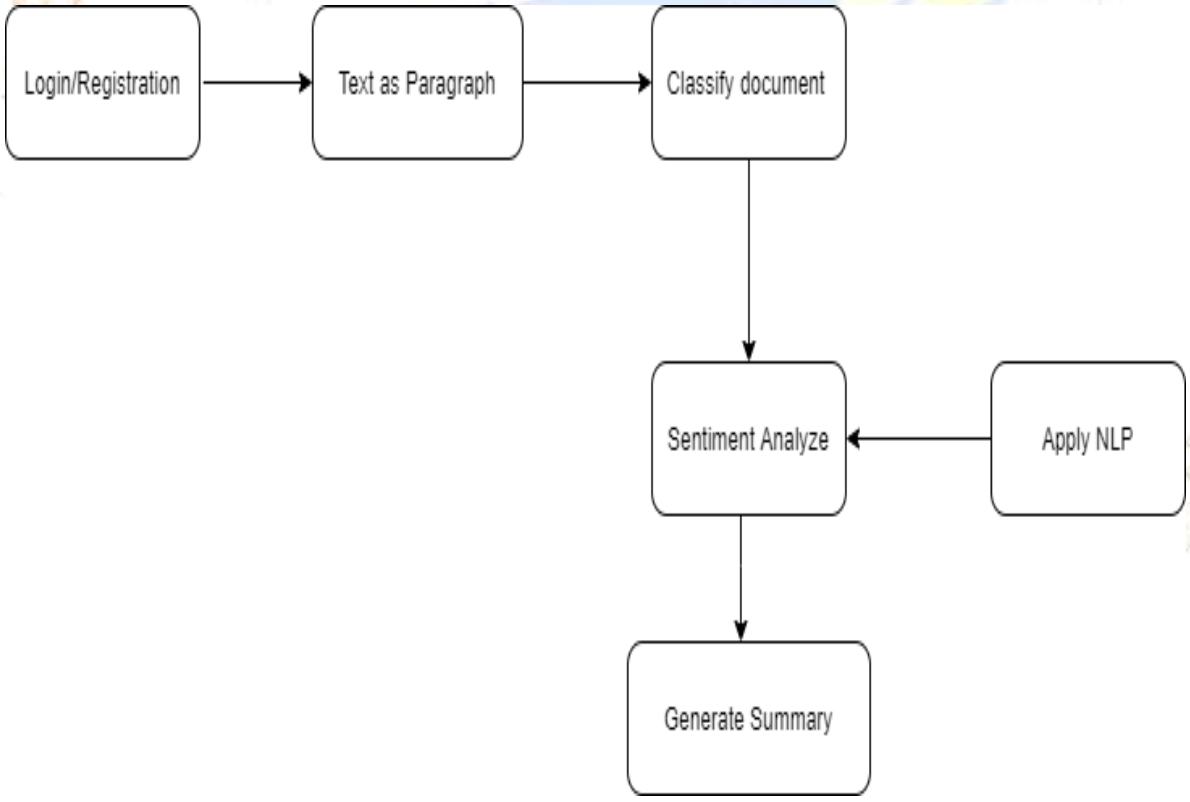

Fig 1: Block Diagram of the Proposed System's Overall Methodology

6. Conclusion

In this study, decision tree, random forest, kernel SVM, ANN, and k-NN were used to create extractive text summaries. Additionally, it was discovered which metrics have the greatest impact on the creation of a summary. The best metrics were important words and word frequency, while the random forest model performed the best overall. The findings would help with model selection, metric construction, and accuracy improvement for summarizers in the future. Although a sizable number of implementations were covered by the tested models, the list is not all-inclusive. There are other additional models that could be the subject of future research.

7. Future Scope

Computerized text summarization has a bright future and is anticipated to see substantial developments and applications across a range of industries. The following are some of the important areas of possible growth and progress in the field of text summarization as technology, research, and user demands continue to change:

1. The ability of text summarization systems to comprehend and produce material will improve, moving beyond simple sentence extraction to provide deeper insights, context, and organized summaries.
2. Summarization systems will develop to accommodate many media types as the digital environment becomes more multimodal with the integration of text, images. For these systems, being able to summarize knowledge in several modalities will be a valuable talent.
3. Systems for summarizing information will get better at processing information in real-time, giving users access to timely summaries whenever they need them. This will be especially useful for monitoring social media, live events, and news updates.
4. There will be an increase in specialized summarizing models that provide highly precise and context-aware summaries customized to specific areas, such as healthcare, finance, and law.
5. Through the simplification of difficult instructional materials, the creation of study aids, and aiding students in understanding essential concepts, summarization systems will play a significant role in e-learning**.**

8. References

[1] R. Socher, “Boiling the information ocean,” 2020.[Online]. Available: http://tiny.cc/45ohlz

[2] S. Chopra, M. Auli, and A. M. Rush, “Abstractive sentence summarization with attentive recurrent neural networks,” in Proc. Conf. North Amer. Chapter Assoc. Comput.Linguist.: Human Lang. Technol., 2016, pp. 93–98.

[3] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence RNNs and beyond,” in Proc. SIGNLL Conf. Comput. Natural Lang. Learn. Stroudsburg, PA, USA, 2016, pp. 280–290.

[4] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in Proc. Annu. Meet. Assoc. Comput Linguist. 2017, pp. 1073–1083.

[5] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” in Proc. Int. Conf. Learn Representations, 2018.

[6] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Language Process. 2019, pp. 3721–3731.

[7] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, “MASS: Masked sequence to sequence pre-training for language generation,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5926–5936.

[8] L. Dong et al., “Unified language model pre-training for natural language understanding and generation,” inAdv. Neural Inform. Process Syst. 2019, pp. 13 042–13 054.

[9] Y. Yan et al., “ProphetNet: Predicting future N-gram for sequence-to sequence pre-training,” 2020, arXiv:2001.04063.

[10] Mr. Vijay Sonawane, Prof. Amit Mishra, Dr. Shiv Sahu “A Modified Technique for Modern Multi-Document Summarization,” in IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 09, 2014 | ISSN (online): 2321-0613.

[11] Mr. Vijay Sonawane, Prof. Rakesh Salam “Graph Based Approach For Multi Document Summarization,” in International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 4 Issue 4 April 2015, Page No. 11134-11138.

[12] Mr. Vijay Sonawane, Amit Mishra, Dr. Shiv Sahu “A Literature Review of Modern Multi-Document Summarization Techniques,” in IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 04, 2014 | ISSN (online): 2321-0613.

[13] Mr. Vijay Sonawane, Rakesh Salam “Result Evaluation of Graph Based Multi Document Summarization,” in International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438