International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2665 - 2670
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0395

Text Summarization Using NLP

Praveen Kumar Gupta[1], Gaurav Dubey[2], Akshat Sharma[3]
[1]Professor, Computer Science Engineering, Galgotias University, Greater Noida, India.
[2,3]Student scholar, Computer Science Engineering, Galgotias University, Greater Noida, India.
Emails: *praveen.gupta@galgotiasuniversity.edu.in[1], gauravdubey2828@gmail.com[2], akshatsharma.1153@gmail.com[3]*

Abstract
This paper provides a detailed view of how the development and usage of a natural language processing text summarizer have occurred. Based on its requirement to present condensed yet meaningful conclusions without affecting the original meaning of the given data, this system is built up. Techniques utilized, like feature extraction, preprocessing of the given data, and summarization itself, have also been mentioned within this research paper. It also reviews how well the system performs in benchmark datasets, and it expounds on some of its applications, limitations, and possible further developments. In this study, it shows quite well how hard-core NLP algorithms can treat the challenges encountered in modern text summarization.

Keywords: *Text summarisation, Information retrieval, Electronic text, Extraction, Abstraction, Online shopping.*

1. Introduction
Because of the exponential proliferation of digital data, it currently mostly specializes in summarizing huge quantities of information into concise yet informative texts. Interactions and understandings of text content are changing because of natural language processing. Summarization of text is perhaps the most widely used application of natural language processing (NLP) to overcome the issue of compacting long texts, articles, or sections into bite-sized pieces without compromising essential information and ideas. Summarization has been practiced for millennia, beginning with human edited summaries over time and continuing through contemporary algorithms driven by computational linguistics. Nonetheless, an automated, efficient text summarizing process is now in greater demand as a result of the requirement to produce content for the Internet and instant information access.

- **Extractive Summarization:** This will enable it to come up with a coherent summary by picking sentences directly from the book and reorganizing them without losing the focus on the most educational sentences. Reading the material, scrubbing it, and computing the sentence scores by employing the Heapq model's n largest () function are all involved in extractive text summarizing. The process of text extraction summarization is as follows:
- Getting the Text Ready
- Evaluation of Sentences
- Choose the Sentences
- Post processing
- **Abstractive Summarization:** By generating new sentences with the identical meaning as the original content, abstractive summarization takes an additional step. This type of method, which often employs approaches such as language generation models, requires a deeper understanding of language and context. Text summarization comes with a range of challenges. In the generation of summaries, NLP models must be able to comprehend the meaning of phrases, consider contextual hints, and ensure grammatical correctness. They must also cope with different formats of text, such as research articles, news reports, social media posts, and others. NLP-based text summarization has numerous practical

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2665 - 2670
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0395

applications. News digests can be automated by news agencies, so readers can quickly grasp what's been happening in the world today. Scientists can easily find relevant studies by browsing through numerous academic articles.

2. Literature Review

Rule-Based Methods: The majority of the early methods of text summarisation used pre-defined heuristics and rules, like keyword frequency and sentence location. Simple and efficient to compute as they are, they tend not to capture the finer context and lack the robustness to accommodate varying text patterns. Jugran et al. [1] employed Spacy in Python to develop extractive text summarisation, illustrating how rule-based NLP pipelines can effectively pick out significant sentences from documents. To make text summarisation systems available to both practitioners and learners, Awasthi et al. [2] presented an in-depth tutorial on how to build them on the basis of basic NLP principles. Recent advances in NLP-based text summarisation, both extractive and abstractive, were comprehensively explored in the Awasthi et al. survey. Prakash et al. [3] implemented an independent summarisation system using Spacy and evaluated its performance on large textual datasets in the medical domain.

2.1.Statistical Methods

Methods such as Latent Semantic Analysis (LSA) and Term Frequency-Inverse Document Frequency (TIDF) provided statistical models for identifying important content. These methods surpassed rule-based methods by detecting statistical relationships in the text, but they were still not able to detect the deeper semantics of the text. To enhance the phrase selection relevance, Khan et al. [4] proposed a hybrid extractive summarisation that incorporates the TF-IDF and K-means clustering. Shetty and Kallimani [5] applied K-means to cluster similar sentences and find the most representative ones to focus on clustering-based summarisation. From Young [6], the goals of summarisation find indirect support by the technical writing rules of structure, conciseness, and clarity. One of the earliest evaluations of the literature that was focused entirely on NLP based single-document summarisation techniques was given by Haque et al. [7].

- **Graph-Based Models:** Text Rank and other graph based models were initially used to rank sentences based on their relevance. Because of their capacity to identify globally important sentences in a set of texts, these methods became popular. Neural Networks and **Deep Learning:** Deep learning revolutionized text summarization with architectures like sequence-to-sequence models with attention mechanisms that are designed to capture longrange dependencies better. **Transformer-Based Models:** Pre-trained transformer models like BERT, GPT, and T5 have set new standards for text summarization. These models employ advanced architectures and massive amounts of training data to achieve state-of-the-art performance in extractive and abstractive summarization tasks.
- **Hybrid Approaches**: Current research focuses on the fusion of extractive and abstractive methods to leverage their strengths. To boost productivity. For instance, abstractive methods can prefilter important sentences for abstractive paraphrasing besides coherence. In their study of summarisation techniques for product reviews, Boorugu and Ramesh [8] found domain specific challenges with sentiment-laden text. Adhikari [9] explored machine learning-based methods for summarisation of text by reviewing both modern neural networks as well as traditional algorithms Christian et al. [10] developed a TFIDF-based extractive summariser and further evaluated how well it summarized individual documents automatically
- **Evaluation Metrics:** Standardized evaluation has been facilitated by widely used metrics like ROUGE and BLEU. But their emphasis on lexical overlap over semantic accuracy makes unnecessary characters, punctuation, and stop-words.

3. Problem Identification

- **Overwhelming Information:** The growing amount of digital text makes it more and more impractical for individuals and organizations to manually summarize material. Most datasets are now underutilized because existing solutions cannot handle scalability.
- **Preserving Context:** Most summarizing

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2665 - 2670
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0395

- methods struggle to preserve the original meaning and context of the content, especially in abstractive summarization. Making accurate and concise summaries is still a major challenge.
- **Grammatical and Semantic Coherence:** Maintenance of grammaticality and semantic coherence in summarized versions can prove to be challenging, especially in abstractive techniques. Uncertainty in the original text could make summaries inexact or unreliable.
- **Maintaining Context:** Most summarizing strategies fail to keep the original context and intent of the content, especially in abstractive summarization.
- **Real-World Applications Analysis:** Demonstrate the utility of summarizers across various real world situations, such as journalism, customer support, and scholarly research.
- **Domain-Specific Challenges:** Specialized texts in domains like law or medicine demand specialized methods so as to appropriately summarize intricate structures and domain-specific terminology. Under such conditions, general-purpose models tend to perform poorly because of the absence of train ing data specific to a domain.

4. Objectives

Build a Robust Text Summarization System: Develop and implement a system integrating extractive and abstractive techniques to generate summaries of high quality. Use Stateof-the-Art NLP Techniques. For more effective summarization, leverage transformer based models and other contemporary NLP methods.

- **Maintain Context and Coherence:** Keep the final summaries as close as possible to the original context, content, and grammatical correctness of the source.
- **Handle Domain-Specific Problems:** Optimizing the system for specific domains such as academic, legal, and medical texts can be achieved by finetuning pre-trained models.
- **Enhance Computational Efficiency:** Explore how to make computers reasonably priced and scalable so that they can be more accessible.
- **Compare Performance in Detail:** In order to ascertain the effectiveness of the system, qualitative assessments (such as human rating) need to be carried out alongside quantitative indicators (such as ROUGE and BLEU).
- **Use Innovative NLP Techniques:** For improved summarisation, employ transformer-based models and other new NLP methods. Maintain Context and Coherence: Ensure that the end summaries retain the original context, content, and grammatical integrity of the source text.
- **Information Overload:** As a result of the exponentially growing amount of digital content, it is becoming increasingly hard for individuals and organizations to summarize manually.
- **Analysis of Real-World Applications:** Highlight the worth of summarisers in a range of real-life situations, such as journalism, customer service, and academic search.

5. Methodology of the System

The following are steps in the extractive text summarizing process: The process of Extractive text summarization includes the following:

- Text Pre-Processing
- Sentence Scoring
- Sentence Selection
- Post-processing

6. Pre-processing of Text Scoring Post-Processing

Pre-processing of text is the act of taking the input text from the user and starting to clean it up by removing them the target of frequent criticism, which in turn increases interest in human evaluation and embedding-based measures. The pre-processing step involves the following actions: Formatting Text At this stage, all the text should be in lowercase. The ease of finding and summarizing important information is facilitated by the standardization of the content. Text Cleaning At this step, unnecessary aspects like spelling errors, punctuation, and special characters are stripped to provide cleaner input data for summarization. Tokenization Tokenisation facilitates processing by splitting the text into small units: Sentence tokenisation involves splitting a paragraph into individual sentences. Following

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2665 - 2670
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0395

sentence tokenisation, word tokenisation splits the phrases further into single words.

- **Sentence Evaluation:** This phase determines the relevance of every sentence in terms of its importance compared to the rest of the text.
- **Choosing Sentences:** It only selects sentences that are highly relevant, assigned a high level of priority. Sentence length is also adjusted in this phase to be shorter.
- **Post- Processing:** This final phase merges sentences with the same meaning and cleans up grammatical errors to produce a clean, coherent summary.

7. Overview of Technologies

- **The Spacy Library:** It is a Python library for natural language processing, i.e., NLP. It was chosen as it is useful for making apps that can work with humongous chunks of text.
- **Heapq:** A Python module named "Heapq" provides an implementation of the heap queue algorithm. It is used for efficient management of a priority queue data structure and is part of Python's built-in library. Implementation of the heap queue algorithm. It is used for efficient management of a priority queue data structure and is part of Python's built-in library.
- **NLP:** A computer program which can understand human language is referred to as natural language processing (NLP). It is a branch of artificial intelligence (AI). As NLP encompasses the two major phases, we have selected it.
- **NLTK:** It is a Python NLP library. Since NLTK provides text processing packages for operations such as tokenization, parsing, and classification, we have chosen it.
- **The flask:** It is a light-weight web framework for Python that assists in building web applications in an easy and straight forward manner. After its invention by Armin Ronacher in 2010, it has become one of the leading Python web development frameworks in use today.

8. Implementation

Tokenizes the input text into words and sentences by sending it through spacy. Put in the data in "doc." Counts the occurrences of each word within the text following the removal of the "stopwords" and the punctuation. Scales the word frequencies by dividing them by the highest frequency. It is tokenized into sentences. It calculates the score of every sentence by adding the normalized frequency of each word that makes up the sentence. Generates the summary by selecting a set of sentences with the highest ratings. Forms the final abstract through the combination of the chosen sentences into a string. Prints the calculated summary and the original text, and their respective lengths in terms of numbers. Joins the selected sentences into a single string to generate the final summary. Prints both the original text and the generated summary, along with their respective lengths in terms of the number of words.

Results

The text summarization website developed using Flask provides users with a convenient platform to summarize text inputs and view both the summarized and original texts. The website consists of two pages: the first page features a textbox where users can input text, and upon submitting, they are redirected to the second page where the summarized and original texts are displayed along with the respective word counts. A major achievement in offering consumers a useful tool for summarizing textual content has been made through the creation and implementation of the text summarization website using Flask. The application of Flask, a lightweight and versatile web framework, enables the website to handle user requests and responses effectively, making the website more responsive and efficient. A major achievement in offering consumers a useful tool for summarizing textual content has been made through the creation and implementation of the text summarization website using Flask. The application of Flask, a lightweight and versatile web framework, enables the website to handle user requests and responses effectively, making the website more responsive and efficient.

Outcome

The users can just summarize text inputs and check the summary and original text on the website built using Flask. The users can enter text in a text box on the site's home page, and on submitting, they are redirected to the second page where they can see the original and summary text together with the word count.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2665 - 2670
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0395

Conclusion

The abstract generation for documents of significant or complex text material is the primary objective of this project. Since documents typically contain extraneous material that can mislead their intended meaning, readers might not be able to obtain the accurate information they are looking for. This technique summarizes the content without altering its meaning by removing unnecessary paragraphs and sentences from long texts with the help of NLP-based text summarization. The tool is designed to be useful to users like students, researchers, and journalists. The students will use it to learn better about subjects, and researchers will use it to read research. Journalists can condense studies and report s in an attempt to give a better presentation. The two types of summarizing are abstractive and extractive. Extractive summary involves the use of important words or portions from the original work in an attempt to preserve the article's meaning. Abstractive summarization will yield a summary of the work. We apply extractive summarization since it tries to maintain words and the article's content. It also maintains the article's exact meaning. The text's score meaning will not change. The extractive text summarization will apply tokenization, sentence scoring, and selection. Tokenization divides words and sentences into tokens so that it can identify common words and put them in the right place.

Future Scope

NLP text summarization finds numerous industrial applications. It is utilized by researchers to post quick summaries of headlines that are quick to read and write for news writers. It would be convenient for researchers to analyze and understand thick scientific texts if they utilized this tool. In general, it can be utilized in a chatbot environment to give relevant and correct replies for a user. Text summary is also an important aspect of marketing because it enables customers to easily understand crucial information regarding a product from concise product reviews and descriptions besides that, as social media marketing goes on, summary can give precise and concentrated information regarding businesses and products. In short, a text summarizer nicely summarizes huge volumes of information and gives its essential meaning. Students, journalists, researchers, and others are the most useful consumers. This text summarization makes content more readable, comprehensible, and concise because most publications contain much unnecessary information.

References

[1]. JUGRAN, S., KUMAR, A., TYAGI, B.S. and ANAND, V., 2021, March. Extractive automatic text summarization using SpaCy in Python & NLP. In 2021 International conference on advance computing and innovative technologies in engineering (ICACITE) (pp. 582-585). IEEE.

[2]. NLP Tutorial 12 - Text Summarization using NLP.

[3]. Prakash, N.C., Narasimhaiah, A.P., Nagaraj, J.B., Pareek, P.K., Maruthikumar, N.B. and Manjunath, R.I., 2022. Implementation of NLP based automatic text summarization using spacy. International Journal of Health Sciences, 6, pp.7508-7521

[4]. Khan, R., Qian, Y. and Naeem, S., 2019. Extractive based text summarization using k-means and tf-idf. International Journal of Information Engineering and Electronic Business, 10(3), p.33.

[5]. Shetty, K. and Kallimani, J.S., 2017, December. Automatic extractive text summarization using K-means clustering. In 2017 international conference on electrical, electronics, communication, computer, and optimization techniques (iceeccot) (pp. 1-9). IEEE.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6]. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7]. Haque, M.M., Pervin, S. and Begum, Z., 2013. Literature review of automatic single document text summarization using NLP. International Journal of Innovation and Applied Studies, 3(3), pp.857 865.

[8]. Boorugu, R. and Ramesh, G., 2020, July. A survey on NLP based text summarization for summarizing product reviews. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 352-356). IEEE.

[9]. Adhikari, S., 2020, March. Nlp based machine learning approaches for text

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2665 - 2670
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0395

summarization. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 535538). IEEE.

[10]. Christian, H., Agus, M.P. and Suhartono, D., 2016. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications, 7(4), pp.285-294.