Text Summarization using NLP

B Rajesh[1], K Nimai Chaitanya[2], P Tejesh Govardhan[3], K Krishna Mahesh[4], B Sudarshan[5]

[1]*Assistant Professor, Dept. of CSE GITAM(Deemed to be University), Visakhapatnam, Andhra Pradesh, India*
[2,3,4,5] *Student, GITAM(Deemed to be University), Visakhapatnam, Andhra Pradesh, India*

---***---

Abstract - *There is a lot of information available on the internet today, but it can be difficult to obtain it quickly and efficiently. It can be difficult to find exactly the information you need to understand a particular topic. The context of the text is useful for extracting the most important information from the various contents of the text. In the age of information overload on the Internet, automatic document summarization is especially important for retrieving important information from many electronic documents. Text summarization that uses natural language processing to summarize text. There are many ways to describe different methods of explanation: extraction and abstraction from single or mixed data; the purpose of content reduction; features of the manuscript; process from shallow to deep; and the content of the article. With the increase in internet and smartphone usage, online shopping has also increased. Everyone wants their belongings to be delivered to their home in good condition. But reading these long reviews is not easy for everyone. Therefore, brevity should be able to reduce long words into short sentences with shorter words describing the same subject.*

Keywords *- Text summarization, information retrieval, electronic text, extraction, abstraction, online shopping.*

1.INTRODUCTION

In an era characterized by the exponential growth of digital information, the ability to distil large volumes of text into concise yet informative summaries have become an essential task. Natural Language Processing (NLP) revolutionizes how we interact with and make sense of textual data. Text summarization, a prominent application of NLP, addresses the challenge of condensing lengthy documents, articles, or passages into shorter versions while retaining the core ideas and critical information.

The art of summarization dates back centuries, from early human-curated abstracts to modern- day algorithms driven by computational linguistics. However, the surge in online content creation and the demand for quick access to information has prompted the need for automated and efficient text summarization techniques. NLP techniques have truly shined, enabling machines to comprehend, analyze, and generate human-like summaries.

There are two primary approaches to text summarization: extractive and abstractive summarization.

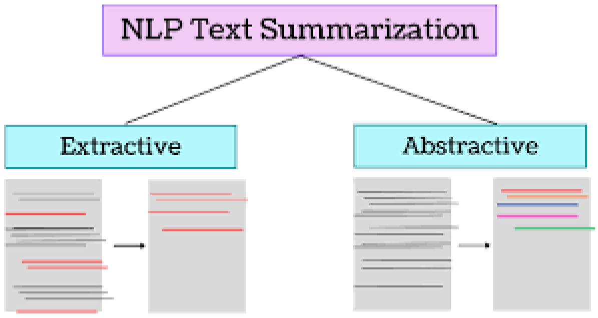


Fig-1: Extractive and Abstractive

1.1 Extractive Summarization

It involves selecting and rearranging sentences directly from the source text and extracting the most informative sentences to form a coherent summary. The methodology of the extractive text summarization includes reading the document, it will then clean the text and then weighs the sentence scores by the nlargest() function in the Heapq model.

The process of Extractive text summarization includes the following:

I. Text Pre-Processing
II. Sentence Scoring
III. Sentence Selection
IV. Post-processing

1.2 Abstractive Summarization

Abstractive summarization goes a step further by generating new sentences that has the same meaning of the original content. This approach requires a deeper understanding of language and context, often involving techniques such as language generation models.

The challenges in text summarization are multifaceted. NLP models must discern the importance of sentences, grasp contextual nuances, and ensure grammatical accuracy while crafting summaries. Additionally, they must handle various text types, such as news articles, research papers, social media posts, and more. The applications of NLP-based text summarization are diverse and impactful. News agencies can automate the process of generating news digests, enabling readers to grasp the day's events quickly. Researchers can sift through numerous academic papers to find relevant studies efficiently. It is used for summarizing the reviews of online products. Content creators can repurpose their articles into shorter formats for different platforms.

2. LITERATURE REVIEW

[1] Shetty, K. et al. Text summarization through clustering and token extraction is a technique that involves breaking down a text into clusters of related content, often using clustering algorithms. Within these clusters, the text is tokenized, resulting in individual words or phrases. Each token is then assigned a relevance score using methods like TF-IDF or Text-Rank. The highest-scoring tokens within each cluster are chosen as representative keywords or phrases. These selected tokens from all clusters are combined to create a summary, offering a concise representation of the primary themes and information found in the original text. This method is valuable for condensing large volumes of text and highlighting key details.

[2] Haque, M.M et al. The increasing volume of internet data has made efficient information retrieval crucial. Users often need help to extract essential information from search results. Automatic text summarization systems offer various benefits, including enhanced research efficiency by eliminating redundant information, improved presentation on small devices, and reduced machine translation time. These advantages make text summarization a valuable tool for managing vast online information.

[3] Boorugu et al. There has been a continuous increase in internet users every year. With increased Internet users comes a great deal of information stored online every second. There is a need to summarize this data while retaining the original meaning of the data. Thus, the process of Text Summarization comes into the picture with its benefits spread over different fields such as Machine Learning, Natural Language Processing, Artificial Learning, and Semantics.

[4] Adhikari, S. et al. In today's data-centric world, where we produce and consume vast amounts of information, text summarization plays a crucial role. It condenses textual content, making it easier to extract essential insights. Various techniques, including machine learning, neural networks, and NLP libraries like scikit-learn and NLTK, are employed to achieve this.

[5] JUGRAN, S. et al. Extractive text summarization is a powerful tool for rapidly extracting the gist of elaborated text. It is particularly useful for scanning large reports and identifying the most important data. This project aims to develop a text summarization model based on the extractive approach, which uses the same set of important words from the original text to generate a shorter, more concise summary. We will evaluate the effectiveness of different methods for distinguishing between different summaries based on size and accuracy.

[6] Christian, H. et al. It discusses the growing need for text summarization due to the rapid increase in textual data, particularly on the internet and social media. It highlights the role of Natural Language Processing (NLP) in text summarization, explaining two methods: extractive and abstractive summarization, with a focus on the Term Frequency-Inverse Document Frequency (TF-IDF) method. It also outlines the four main steps in building an extractive summarization program, from preprocessing to summary generation.

[7] The extractive text summarization involves the steps of text-cleaning, formatting and tokenization. It will collect words that are frequently repeated and collects into word frequency table so that it can know about the meaning and importance of the word.

[8] Awasthi, et al. Text summarization is a NLP challenge where software shortens texts to highlight key points. It's categorized by input type (single/multi-document, generic/domain- specific, query-based) and output type (extractive/abstractive). Applications include media monitoring and more.

[9] Prakash, N.C. et al. Automatic text summarization using spaCy is a process that employs the spaCy library for natural language processing. It begins by parsing and tokenizing the input text and then scores individual sentences based on their relevance. This is typically accomplished using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. The sentences with the highest scores are selected to form a concise summary, effectively condensing lengthy documents or articles into shorter, coherent versions.

[10] Khan, R. et al. Text summarization is a technique to extract the most essential information from a text document. The experimental results show that this method is used to achieve good results for extractive text summarization.

3. Problem Identification

Text summarization is the process of generating a shorter version of a text document without removing the most essential information. This is because different people may have different opinions on what is important, and the importance of information may also depend on the context. Problem with text summarization is that it is difficult to generate a summary that is both fluent and informative Summaries often need to be concise, but they should also be easyto read and understand.

4. OBJECTIVES

The main objective is to develop algorithms that can automatically generate accurate, fluent, and concise summaries of text documents.

Specifically, NLP-based text summarization systems should be able to:

- Generate a summary that is fluent and easy to read.
- Generate a summary that is concise, but still contains the same information from the text.
- Be able to summarize a variety of text types, including news articles, scientific papers, and email messages.

5. SYSTEM METHODOLOGY

The process of Extractive text summarization includes the following:

I. Text Pre-Processing
II. Sentence Scoring
III. Sentence Selection
IV. Post-processing

I. Text Pre-Processing: This step involves taking the input text from the user, and then it starts cleaning the text by removing stop words, punctuation, and other irrelevant characters.

The pre-processing involves three steps of the process:

a) Text formatting is the process of formatting the text into lowercase letters. It will edit all the letters of the text into lowercase. It will be easy to summarize the essential points.
b) Text- Cleaning: It is the process of removing spelling mistakes, punctuation, and special characters present in the text.
c) Tokenization: It is the process of breaking the words into single tokens. It will divide both the words and sentences.
   (1) Sentence Tokenization: The sentence tokenization will break a paragraph into sentences.
   (2) Word Tokenization: The word tokenization will occur after the sentence tokenization, and then the word in the sentence will be divided.

II. Sentence Scoring: It is the process of measuring the importance of the sentence.

III. Sentence Selection: The sentences that get the top priority will be selected. The length of the sentence will also be reduced.

IV. Post-Processing: This step involves merging the sentences with the same meaning and correcting the text's grammatical errors.

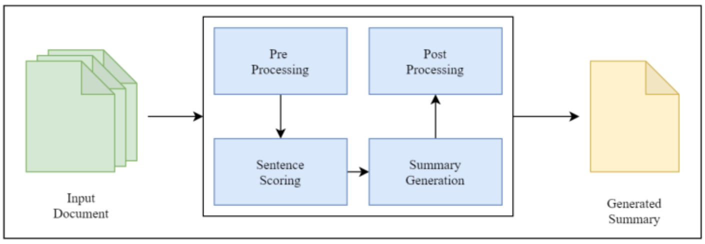


Fig-2: Text-Summarization process

6. OVERVIEW OF TECHNOLOGIES

Spacy library:

It is a library in python used in Natural Language Processing (NLP). We have chosen this library because it helps in building applications that process large volumes of text.

Heapq:

"heapq" is a Python module that provides an implementation of the heap queue algorithm.
It is part of Python's standard library and is used for efficiently maintaining a priority queue data structure.

NLP:

NLP is a computer program, that has the ability to understand human language. It is a component of artificial intelligence (AI). We have chosen NLP because it contains the two main phases.

NLTK:

It is a library in python used in Natural Language Processing (NLP). We have chosen NLTK because it has the libraries of text processing to perform the operations like tokenization, parsing, and classification.

Flask:

It is a lightweight and web framework for Python, used to make web applications quick and easy. It was created by Armin Ronacher in 2010 and has since become one of the most popular frameworks for web development in Python.

7. IMPLEMENTATION

7.1 METHODOLOGY

1. Install the libraries
   i. Setup Tools
   ii. Spacy
   iii. "en_core_web_sm" libraries.

2. Loads the English language model provided by spaCy.

3. Processes the input text using spaCy to tokenize it into words and sentences.

4. Load the input into "doc"

5. Counts the frequency of each word in the text after filtering out "stopwords" and "punctuation" marks.

6. Normalizes the word frequencies by dividing them by the maximum frequency.

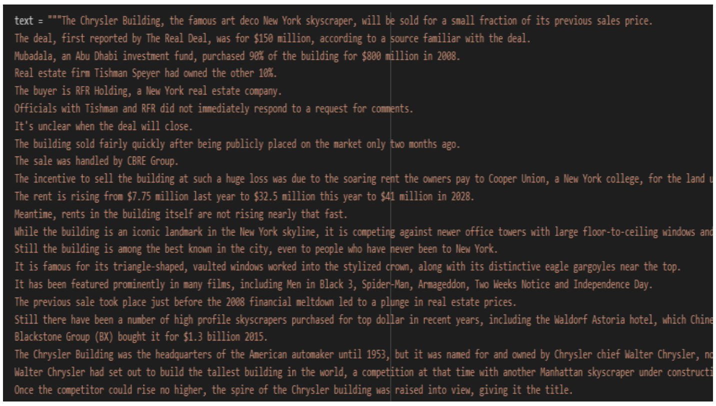
7. Tokenizes the text into sentences.

8. Calculates a score for each sentence based on the sum of normalized frequencies of the words it contains.

9. Selects a subset of sentences with the highest scores to form the summary.

10. Joins the selected sentences into a single string to generate the final summary.

11. Prints both the original text and the generated summary, along with their respective lengths in terms of the number of words.



```
text = """The Chrysler Building, the famous art deco New York skyscraper, will be sold for a small fraction of its previous sales price.
The deal, first reported by The Real Deal, was for $150 million, according to a source familiar with the deal.
Mubadala, an Abu Dhabi investment fund, purchased 90% of the building for $800 million in 2008.
Real estate firm Tishman Speyer had owned the other 10%.
The buyer is RFR Holding, a New York real estate company.
Officials with Tishman and RFR did not immediately respond to a request for comments.
It's unclear when the deal will close.
The building sold fairly quickly after being publicly placed on the market only two months ago.
The sale was handled by CBRE Group.
The incentive to sell the building at such a huge loss was due to the soaring rent the owners pay to Cooper Union, a New York college, for the land u
The rent is rising from $7.75 million last year to $32.5 million this year to $41 million in 2028.
Meantime, rents in the building itself are not rising nearly that fast.
While the building is an iconic landmark in the New York skyline, it is competing against newer office towers with large floor-to-ceiling windows and
Still the building is among the best known in the city, even to people who have never been to New York.
It is famous for its triangle-shaped, vaulted windows worked into the stylized crown, along with its distinctive eagle gargoyles near the top.
It has been featured prominently in many films, including Men in Black 3, Spider-Man, Armageddon, Two Weeks Notice and Independence Day.
The previous sale took place just before the 2008 financial meltdown led to a plunge in real estate prices.
Still there have been a number of high profile skyscrapers purchased for top dollar in recent years, including the Waldorf Astoria hotel, which Chine
Blackstone Group (BX) bought it for $1.3 billion 2015.
The Chrysler Building was the headquarters of the American automaker until 1953, but it was named for and owned by Chrysler chief Walter Chrysler, no
Walter Chrysler had set out to build the tallest building in the world, a competition at that time with another Manhattan skyscraper under constructi
Once the competitor could rise no higher, the spire of the Chrysler building was raised into view, giving it the title.
```

Fig-3: Input Data

7.2 TESTING

Torch:

PyTorch is a powerful and versatile machine learning library in Python, offering dynamic computation graphs, tensor operations, automatic differentiation, and a high-level neural network module. Its flexibility, ease of use, and growing ecosystem make it a preferred choice for both research and production-level deep learning projects.

Rouge:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) stands as a comprehensive suite of metrics designed to assess the efficacy of automatic summarization algorithms in producing text summaries. This evaluation method gauges the degree of overlap between summaries generated by systems and predetermined reference summaries. By analyzing this overlap, ROUGE offers valuable insights into the effectiveness of a summary in encapsulating essential information from the source text.

1. ROUGE-1 Score (Unigram Overlap):

A recall of 0.5795 indicates that 57.95% of the unigrams present in the reference summary were also captured in the generated summary. Similarly, a precision of 0.3835 suggests that 38.35% of the unigrams in the generated summary are relevant to the reference summary. The F1-score of 0.4615 provides a balanced evaluation of recall and precision (harmonic mean) providing a balanced measure of the summarization's accuracy.

2. ROUGE-2 Score (Bigram Overlap):

A recall of 0.4167 indicates that 41.67% in the reference summary were matched in the generated summary, while a precision of 0.2674 suggests that 26.74% in the generated summary are relevant. The F1-score of 0.3257 reflects the balance between recall and precision for bigram overlap.

3. ROUGE-L Score (Longest Common Subsequence):

A recall of 0.5682 suggests that 56.82% of the LCS in the reference summary was captured in the generated summary, while a precision of 0.3759 indicates that 37.59% of the LCS in the generated summary is relevant. The F1-score of 0.4525 provides a balanced evaluation of recall and precision for LCS overlap.

The ROUGE scores provide valuable insights into the accuracy of the text summarization algorithm. Higher the scores, particularly for "ROUGE-1" and "ROUGE-L", indicate a stronger alignment between the generated summary and the reference summary, suggesting a more accurate summarization process.

In conclusion, the provided ROUGE scores offer quantitative measures of the summarization algorithm's performance, indicating its ability to capture key information from the original text effectively.

Table -1: Scores

Name	'r'	'p'	'f'
'rouge-1'	0.579	0.383	0.461
'rouge-2'	0.417	0.267	0.325
'rouge-l'	0.568	0.375	0.452

GRAPHICAL INTERPRETATION:

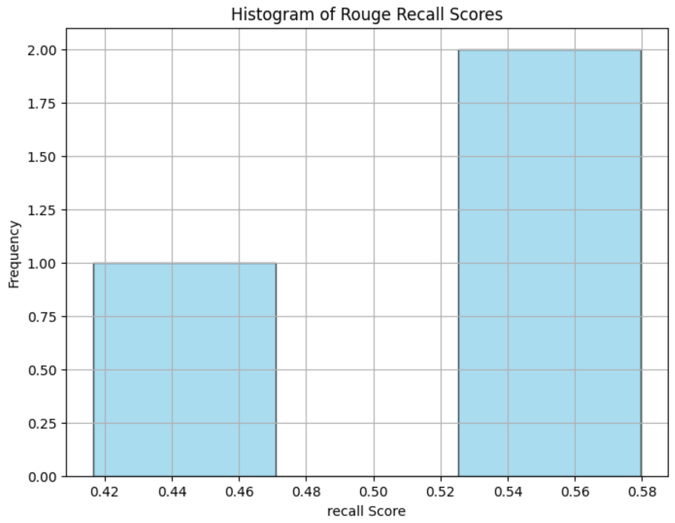


Fig-4: Recall scores

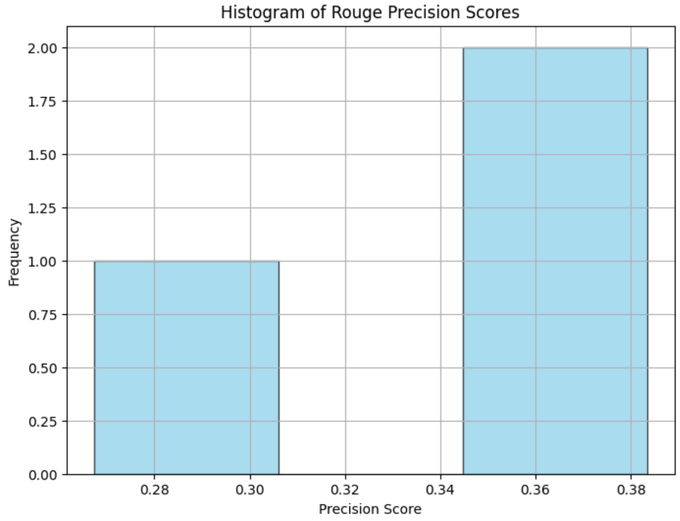


Fig-5: Precision scores

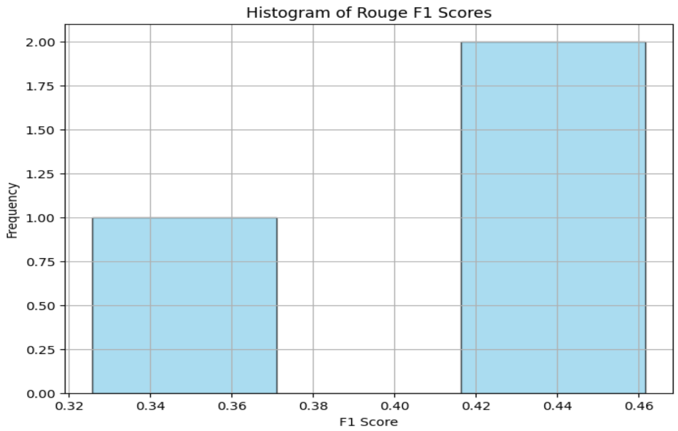


Fig-7: F1 scores

8. RESULTS

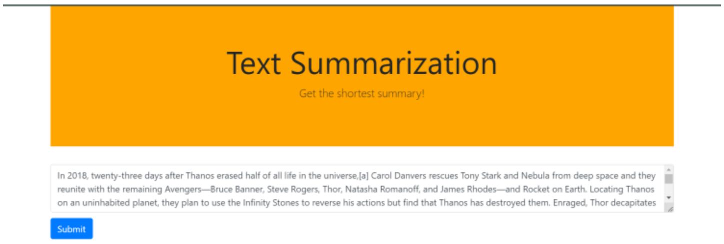
The text summarization website developed using Flask provides users with a convenient platform to summarize text inputs and view both the summarized and original texts. The website consists of two pages: the first page features a textbox where users can input text, and upon submitting, they are redirected to the second page where the summarized and original texts are displayed along with the respective word counts.

Key features and functionalities:

i. Input Interface

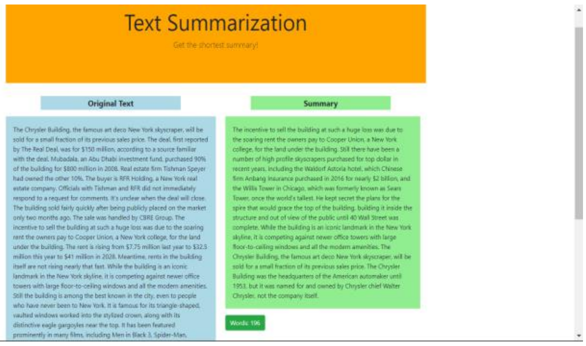
ii. Text Summarization

iii. Word Count Display



In 2018, twenty-three days after Thanos erased half of all life in the universe,[a] Carol Danvers rescues Tony Stark and Nebula from deep space and they reunite with the remaining Avengers—Bruce Banner, Steve Rogers, Thor, Natasha Romanoff, and James Rhodes—and Rocket on Earth. Locating Thanos on an uninhabited planet, they plan to use the Infinity Stones to reverse his actions but find that Thanos has destroyed them. Enraged, Thor decapitates

Fig-8: Input Page

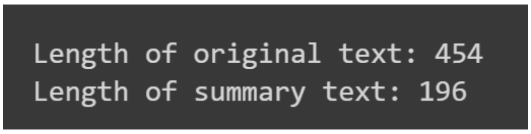


The Chrysler Building, the famous art deco New York skyscraper, will be sold for a small fraction of its previous sales price. The deal, first reported by The Real Deal, was for $150 million, according to a source familiar with the deal. Mubadala, an Abu Dhabi investment fund, purchased 90% of the building for $800 million in 2008. Real estate firm Tishman Speyer had owned the other 10%. The buyer is RFR Holding, a New York real estate company. Officials with Tishman and RFR did not immediately respond to a request for comments. It's unclear when the deal will close. The building sold fairly quickly after being publicly placed on the market only two months ago. The sale was handled by CBRE Group. The incentive to sell the building at such a huge loss was due to the soaring rent the owners pay to Cooper Union, a New York college, for the land under the building. The rent is rising from $7.75 million last year to $32.5 million this year to $41 million in 2028. Meantime, rents in the building itself are not rising nearly that fast. While the building is an iconic landmark in the New York skyline, it is competing against newer office towers with large floor-to-ceiling windows and all the modern amenities. Still the building is among the best known in the city, even to people who have never been to New York. It is famous for its triangle-shaped, vaulted windows worked into the stylized crown, along with its distinctive eagle gargoyles near the top. It has been featured prominently in many films, including Men in Black 3, Spider-Man,

The incentive to sell the building at such a huge loss was due to the soaring rent the owners pay to Cooper Union, a New York college, for the land under the building. Still there have been a number of high profile skyscrapers purchased for top dollar in recent years, including the Waldorf Astoria hotel, which Chinese firm Anbang Insurance purchased in 2016 for nearly $2 billion, and the Willis Tower in Chicago, which was formerly known as Sears Tower, once the world's tallest. He kept secret the plans for the spire that would grace the top of the building, building it inside the structure and out of view of the public until 40 Wall Street was complete. While the building is an iconic landmark in the New York skyline, it is competing against newer office towers with large floor-to-ceiling windows and all the modern amenities. The Chrysler Building, the famous art deco New York skyscraper, will be sold for a small fraction of its previous sales price. The Chrysler Building was the headquarters of the American automaker until 1953, but it was named for and owned by Chrysler chief Walter Chrysler, not the company itself.

Fig-9: Output Page

The development and deployment of the text summarization website using Flask mark a significant milestone in providing users with a practical tool for summarizing textual content. The integration of Flask, a lightweight and flexible web framework, enables efficient handling of user requests and responses, contributing to the website's responsiveness and performance.



```
Length of original text: 454
Length of summary text: 196
```

Fig-10: Words displayed

9. CONCLUSIONS

In conclusion, the project aims to provide a summarization of the articles that contain significant or large text content. Some articles provide unnecessary content that changes the meaning of the article. So, the user may need help to get the exact information a user wants. Here, Text Summarization using NLP plays a crucial role by summarizing the text without changing the exact meaning of the text. It also helps to remove unnecessary sentences and paragraphs from vast text content.

This project will be helpful for students, researchers, and news reporters. They can use this tool to summarize the topic of a research paper or any news report to understand and present it.Students can use this tool to read and understand any topic more thoroughly. Researchers can use this tool to study any research work they are doing. Text summarization can be done extractively and abstractly. Extractive summarization involves the extraction of the text from the enormous text article such that the meaning of the article does not change. In contrast, abstractive summarization will write a summary of the article.

Here, we use extractive summarization because it tries to keep the article's sentences and the meaning the same. It also maintains the exact meaning of the article. It will not change the essential meaning of the text content. The extractive text summarization will involve the process tokenization, sentence scoring and selection. Tokenization divides the sentences and words to tokens so that it can find the word that has been repeatedly frequently so that it places the word in correct place.

10. FUTURE SCOPE

The text summarizing using NLP has much scope in many industries. News reporters can use it to write the latest news headlines with concise overviews. It will be mainly used in the research industry to study and understand complex scientific papers by researchers. Chatbots can also use this text summarization to provide users with precise information. Text summarization is going to be used in marketing. Also, if a person wants to buy a product, the description and review of the product should be precise enough so that product information will be precise. Social media marketing is also increasing nowadays, to provide precise information about a brand and its product, text summarization is useful.

In summary, the text summarizer will provide the summary by keeping the meaning of the vast text content the same. It will be more helpful for students, researchers, news reporters, and general users. People will widely use text summarization because articles nowadays contain unnecessary content. Text summarizer plays a significant role in making it shorter, simpler, and more understandable.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Adhikari, S., 2020, March. Nlp based machine learning approaches for text summarization. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 535-538). IEEE.

[5] 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.

[6] 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.

[7] NLP Tutorial 12 - Text Summarization using NLP

[8] Awasthi, I., Gupta, K., Bhogal, P.S., Anand, S.S. and Soni, P.K., 2021, January.Natural language processing (NLP) based text summarization-a survey. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 1310- 1317). IEEE.

[9] 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

[10] 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.