2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science

ATS: Auto Text Summarization using Natural Language Processing

Ameya Gohad
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India
Ameya.gohad.batch2021@sitnagpur.siu.edu.in

Anay Vyawahare
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India
Anay.vyawahare.batch2021@sitnagpur.siu.edu.in

Anmol Gupta
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India
Anmol.gupta.batch2021@sitnagpur.siu.edu.in

Kashish Sharma
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India
Kashish.sharma.batch2021@sitnagpur.siu.edu.in

Keyur Dhage
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India
Keyur.Dhage.batch2021@sitnagpur.siu.edu.in

Nilesh Shelke
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India
nilesh.shelke@sitnagpur.siu.edu.in

Jagdish Patani
Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International Deemed University, Pune, India

***Abstract*— Automatic Text Summarization (ATS) is a crucial task in Natural Language Processing (NLP), seeking to distill critical information from voluminous textual data. This review provides a concise overview of extractive and abstractive summarization techniques. We explore the methodology used in various papers and extractive features usually implemented in building models like text summarization. Literature review, process, and comparative results are discussed, offering analysis of prominent approaches. This review is a quick reference for researchers and practitioners navigating the evolving landscape of auto-text summarization.**

***Keywords*—Abstractive Summarization, Extractive Summarization, Stop World Removal, Clustering, Tokenization**

I. Introduction

Short writing is a critical part of giving a great speech, which is a significant hassle and supply of subject today. Its main characteristic is to offer a concise summary of the authentic report, maintaining its primary content material. This method makes retrieving records from the database easy and speedy. Without the content of the textual content, humans will have to scan all of the data to get the crucial content material, inflicting humans to waste plenty of time. Manually summarizing many files, merchandise, and diverse files is a completely heavy and tiring task. Writing notes can lessen this burden and give readers a first-rate way to keep time without having to cautiously examine the complete text.

The origins of the script date back to the Fifties, and the field has endured to adapt in view that then. The key to this procedure is an extraction method that recycles facts that the machine deems applicable to the context. This approach uses natural language processing and statistical strategies to create content material. Traditional strategies form the idea of many varieties of article writing.

The importance of writing documents covers many areas of business, verbal exchange, records garage, facts mining, and word processing, and can be done properly. The primary purpose of this text is to explore numerous uses of the short textual content, together with such works as film reviews, email content material, newsletters, writing for college students, and writing advice for specialists in enterprise, government, and medicinal drugs.

The foremost cause of this study is to test the quality of the statistics within the writing of written textual content through the prism of language processing.

II. Litrature Review

In the contemporary landscape, text summarization has gained significant attention. The 21st century is often referred to as the Digital Age, marked by the surge in information technology and related advancements. Automatic text summarization addresses the challenge of reducing the length of a document while preserving its essential content [1, 2, 3].

This systematic review delves into various techniques and methodologies employed in text summarization across diverse applications. These encompass Natural Language Processing (NLP), supervised Machine Learning (ML) technique, Neural Networks (NNs), and the K-nearest neighbours (KNN) algorithm. Different algorithms are harnessed, including word vector embedding, k-nearest neighbor, and even human learning-based approaches. The

979-8-3503-4846-0/24/$31.00 ©2024 IEEE

review also involves a comparative analysis of these techniques based on their performance with various datasets.

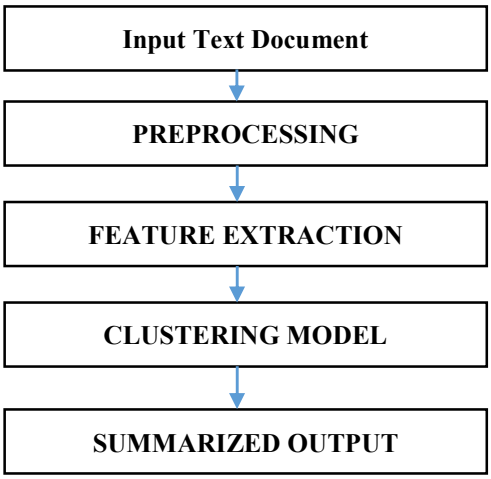

Fig. 1. Natural language processing flow diagram

The realm of text summarization comprises an array of methods, with extractive summarization and abstractive summarization being the most prevalent, as discussed below.

2.1 Extractive Summarization

Extractive summarization crafts a summary by identifying critical words and sentences from the source document, retaining their original context. It selects sentences and words with the highest significance to compile the summary. Statistical features such as document titles, term frequency, and word location are utilized to score and identify the most pertinent content. While extractive summarization is straightforward, it may occasionally introduce errors, such as miscommunication and textual ambiguities in the summary [4,5,6,7].

Fig.2. Extractive Summarization

2.2 Abstractive Summarization

Abstractive summarization takes a different approach by focusing on comprehending the document and then generating a summary that encapsulates its primary message. It mirrors the way humans summarize text. Abstractive text summarization predominantly relies on linguistic principles, aiming to establish semantic connections between words and emphasize the central theme of the document. This approach is more intricate to implement but yields a more human-like summary with reduced uncertainties. Advanced heuristic algorithms are applied in abstractive summarization to minimize redundancy in the summary. The data processing steps encompass the removal of superfluous sentences and words, along with tokenization, to accentuate the central theme of the original document [8,9,10,11].

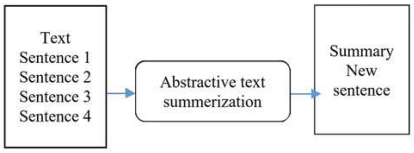

Fig.3. Abstractive Summarization

III. METHODOLOGY

Auto summarization is a challenging task in the field of Natural Language Processing (NLP) that aims to condense a longer text into a shorter, coherent summary while preserving its essential information. Here's a comprehensive methodology for building an auto summarization system [12,13].

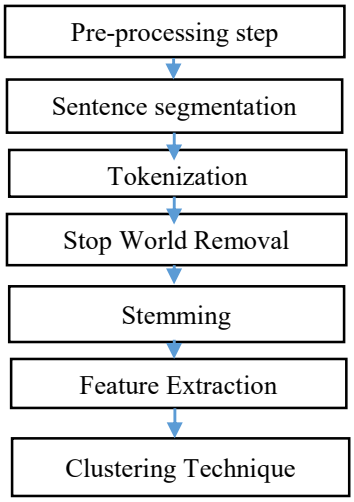

Fig. 4. Proposed Methodology Flow

A. Preprocessing Step

In natural language processing (NLP), preprocessing involves a series of steps applied to raw text data before feeding it into machine learning models or other NLP algorithms. Tokenization breaks down the text into smaller units like words or sub words, providing a foundation for analysis. Lowercasing ensures uniformity by converting all text to lowercase. Stop words, common words lacking substantial meaning, are often removed to reduce noise. Punctuation and special characters are typically eliminated to simplify the text. Stemming and lemmatization help reduce words to their base or root form. Handling numerical values and missing data requires thoughtful consideration. Removing HTML tags or URLs may be necessary for data from web sources. Finally, spell-checking and correction can be applied to standardize the text. The goal is to prepare the text in a format conducive to meaningful analysis and modelling [14,15,16,17,18].

B. Feature Extraction

Text summarization is the process of shortening a text while preserving its essential information. Here are some common

methods used in Natural Language Processing (NLP) for feature extraction in text summarization:

1. Term Frequency – Inverse Document Frequency (TF-IDF):
   - TF-IDF is a statistical measure that indicates the significance of a word to a document in a corpus. It's often used to assign weights to terms in a document.
   - Words with high TF-IDF scores are deemed important and may be included in the summary [19].
2. Sentence Importance Measures: - The importance of each sentence is calculation is based on various features such as word frequency, sentence length, and position within the document [20].
   - For instance, the first and last sentences of a document may be deemed more important.
3. Sentence Position:
   - Greater importance is assigned to sentences at the beginning and end of a document. These positions often contain key information [20].
4. Named Entity Recognition (NER):
   - Named entities (e.g. people, locations, organizations) are identified and extracted from the text [21].
   - Sentences containing important named entities may be considered more relevant for the summary.
5. Word Embeddings:
   - Words are represented as dense vectors in a continuous vector space. Word embeddings like Word2Vec, GloVe, or FastText can capture semantic relationships between words.
   - The similarity between word vectors can be used identify important words and phrases [22,23,24].
6. Sentence Embeddings:
   - Sentences are represented as vectors using techniques like Doc2Vec or Universal Sentence Encoder.
   - Similar to word embeddings, sentence embeddings capture semantic information and can be used to measure sentence similarity.
7. Text Rank Algorithm:
   - Graph-based algorithms are applied to represent the text as a graph, where sentences are nodes are relationships between them are edges [25,26].
   - The importance of each sentence is calculated based on graph centrality measures such as PageRank.
8. Latent Semantic Analysis (LSA):
   - Dimensionality reduction techniques like singular value decomposition are applied to identify latent semantic structures in the text.
   - Sentences with high importance in terms of latent semantic content may be included in the summary.
9. Sentence Compression:
   - Important phrases or sub-sentential units that convey essential information are identified and extracted.
   - Techniques like submodular optimization can be used to select a subset of informative sentences [27].
10. Length and Redundancy Constraints:
   - The length of the summary is limited to predefined number of words or sentences.
   - Redundancy is avoided by penalizing the inclusion of information that has already been covered in the summary.

These features and techniques can be mixed and matched in various ways depending on the specific requirements and goals of the summarization task. The choice of feature extraction methods often depends on the type of summarization (extractive or abstractive) and the characteristics of the input text.

C. Clustering Techiniques

Clustering is a machine learning technique that groups or clusters data by grouping similar items based on unrepealable characteristics in order to reveal patterns and structures within the dataset.

D. Summary Generation

The streamlined processes of creating a shortened, coherent representation of a document's main content while preserving key information for efficient liaison and wringing is known as summary generation.

IV. COMPARITIVE RESULTS

The objective of our experiments was to evaluate the effectiveness of various centrality measures in pinpointing the most crucial aspects of the text. As a result, we took into account measures such as pagerank, hits, closness, betweenness, and degree.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics designed to evaluate the quality of machine-generated text summaries by comparing them to reference (human-generated) summaries. ROUGE-N, where N represents the size of n-grams (word sequences), includes metrics like ROUGE-1 (unigram), ROUGE-2 (bigram), and so on, measuring the overlap of words and phrases between generated and reference summaries. ROUGE-L (Longest Common Subsequence) assesses the longest common sequence of words in the same order, providing insight into the structural similarity of summaries. These metrics are essential in natural language processing for objectively gauging the effectiveness of automatic text summarization systems.

A variety of techniques have been evaluated using two distinct metrics: sentence overlap and sentence edit distance (TED). Following the application of strategies outlined in the methods section, the top five sentences, a quantity determined by the average summary length, are selected to represent the summary generated by our approach(s). This summary is then contrasted with the original summary included in the dataset using the ROGUE metric.

TABLE 1: F score values by the sentence overlap metric and T = 0.5

Method	Rogue-1	Rogue-2	Rogue-L
Pagerank	**0.21**	**0.07**	**0.243**
Hits	**0.23**	0.08	0.244
Closeness	**0.24**	0.08	0.24
Betweenness	**0.23**	0.08	0.24
Degree	**0.239**	0.09	0.24
Clusters	**0.24**	0.07	0.24

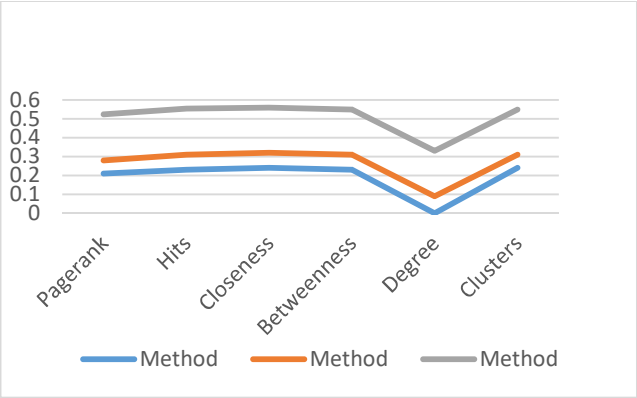

Fig. 5.Graph for F score values by the sentence overlap metric and T = 0.5

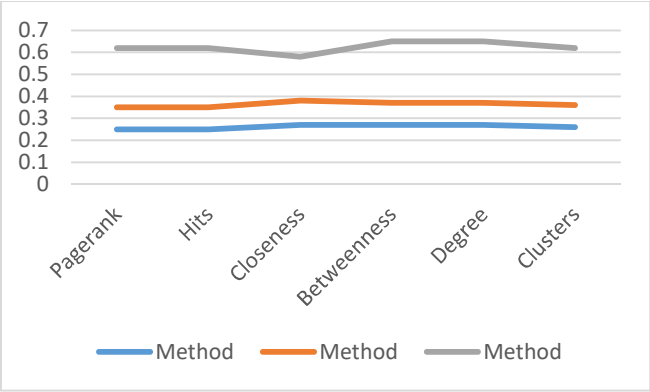

Fig. 6. Graph for F score using TED metric and T = 0.5

TABLE. 2: F score using TED metric and T = 0.5

Method	Rogue-1	Rogue-2	Rogue-2
Pagerank	0.25	0.1	0.27
Hits	0.25	0.1	0.27
Closeness	0.27	0.11	0.2
Betweenness	0.27	0.1	0.28
Degree	0.27	0.1	0.28
Clusters	0.26	0.1	0.26

From the table 1, fig 5 and table 2, and fig. 6 it can be inferred that Closeness and Betweenness works better than other methods.

V. Conclusion

This paper concludes by offering a thorough analysis of automatic text summarizing methods within the field of natural language processing (NLP). We have examined the transition from traditional techniques to modern strategies, evaluating the benefits and drawbacks of abstractive and extractive summarization. Important elements are examined, including feature extraction and sentence significance metrics. The benchmarking of summarization systems was clarified by discussing evaluation measures, such as ROUGE, and it was discovered that closeness and betweenness outperforms other techniques.

References

[1] Chetana Varangatham, J.Srijana Reddy, Uday Yellen, Madhumitha Kotha and Dr. P Venkateswara Rao “Text summarization using NLP”, Journal of emerging technology and innovative research (JETIR), Volume 9, issue 5, may 2022.

[2] Sheetal Patil, Avinash Pawar, Siddhi Khanna, Anurag Tiwari and Somay Trivedi “Text Summarizer using NLP (Natural Language Processing)”, Journal of Computer Technology & Applications, Volume 12, Issue 3, 2021.

[3] Subas Voleti, Chaitan Raju, Teja Rani and Mugada Swetha “Text summarization using natural language processing and google text to speech api”, International Research Journal of Engineering and Technology (IRJET) Volume 7, Issue 5, May 2020.

[4] Rabia Tehseen, Uzma Omer, Muhammad Shoaib Farooq and Faiqa Adnan “Text Summarization Techniques Using Natural Language Processing: A Systematic Literature Review”, Volume 9, Number 4, October-December 2021.

[5] Aaksh Srivastav, Kamal Chauhan, Himanshu Daharwal, Nikhil Mukati and Pranoti Shrikant Kavimandan ”Text Summarizer Using NLP (Natural Language Processing)”, IRE Journals, Volume 6, Issue 1, JUL 2022.

[6] Dalwadi, Bijal Patel, Nikita, and Suthar Sanket, “A Review Paper on Text Summarization for Indian Languages” International Journal for Scientific Research Development (IJSRD), Vol. 5, Issue 07, 2017.

[7] Chandra Khatri, Sumanvoleti, Sathish Veeraraghavan, Nish Parikh, Atiq Islam, Shifa Mahmood, Neeraj Garg, and Vivek Singh, “Algorithmic Content Generation for Products”. *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, USA, 2015, pp. 2945-2947, doi: 10.1109/BigData.2015.7364131.

[8] Obaid, I., Farooq, M. S., Abid, A., "Gamification for recruitment and job training: model, taxonomy, and challenges", IEEE Access, 65164-65178, (2020).

[9] Deepali K. Gaikwad and C. Namrata Mahender, "A Review Paper on Text Summarization”. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[10] Batra, S. Chaudhary, K. Bhatt, S. Varshney and S.Verma, "A Review: Abstractive Text Summarization Techniques using NLP," 2020 International Conference on Advances in Computing, Communication Materials (ICACCM), Dehradun, India, pp. 23-28, doi: 10.1109/ICACCM50413.2020.9213079, 2020.

[11] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut, “Text Summarization Techniques: A Brief Survey”. In Proceedings of arxiv, USA, July 2017.

[12] Thomaidou, I. Lourentzou, P. Katsivelis-Perakis, and M. Vazirgiannis, "Automated Snippet Generation for Online Advertising", Proceedings of ACM International

Conference on Information and Knowledge Management (CIKM'13), San Francisco, pp.1841- 1844, USA, 2013.
[13] Huong Thanh Le and Tien Manh Le, "An approach to Abstractive Text Summarization", In proceeding of International Conference of Soft Computing and Pattern Recognition (SoCPaR), Hanoi, Vietnam, Dec 2013.
[14] Adhika Widyassari, S. R. " Review of automatic text summarization techniques & methods", Journal of King Saud University - Computer and Information Sciences, 18, 2020.
[15] Okumura, H. T. "Text Summarization Model based on the budgeted median problem. Proc. 18th ACM Conf. Inf. Knowledge", 1-4, 2009.
[16] Sutskever, Ilya Vinyals, Oriol and Le, Quoc, "Sequence to Sequence Learning with Neural Networks", Advances in Neural Information Processing Systems, 2014.
[17] Mr. Vikrant Gupta, M. P, "An Statiscal Tool for Multi-Document Summarization", International Journal of Scientific and Research (ISSN 2250-3153), (2012).
[18] Neelima Bhatia, A. J, "Literature Review on Automatic Text Summarization: Single and Multiple Summarizations", International Journal of Computer Applications, 1-5, (2015).
[19] Yllias Chali, S. A, "Query-focused multi-document summarization: automatic data annotations and supervised learning approaches", Cambirdge University Press, (2011).
[20] Suleiman, Dima and Arafat Awajan ,"Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures and challenges", Mathematical Problems in Engineering 2020 (2020).
[21] Jo, "K nearest neighbor for text summarization using feature similarity", International Conference on Communicat, (2017).
[22] Ordonez, Y. Zhang and S. L. Johnsson, "Scalable machine learning computing a data summarization matrix with a parallel array DBMS," Distrib. Parallel Databases, vol. 37, no. 3, pp. 329–350, doi: 10.1007/s10619-018-7229-1, Sep. 2019.
[23] Adhikari, Surabhi. "Nlp based machine learning approaches for text summarization." 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2020.
[24] Verma, Pradeepika, and Anshul Verma, "A review on text summarization techniques", Journal of Scientific Research 64.1 (2020).
[25] Tehseen, R., Farooq, M. S. & Abid, A. ,"Earthquake prediction using expert systems: a systematic mapping study", Sustainability, 12(6), 2420, (2020).
[26] Farooq, M. S., Riaz, S., Abid, A., Umer, T. & Zikria, Y. B. , "Role of IoT technology in agriculture: A systematic literature review", Electronics, *9*(2), 319, (2020).
[27] Farooq, M. S., Khan, S. A., Abid, K., Ahmad, F., Naeem, Shafiqa & Abid, A., "Taxonomy and design considerations for comments in programming languages: a quality perspective", Journal of Quality and Technology Management, 10(2), 167-182, (2015).