IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Comparative Analysis Of Extractive Text Summarization Approaches

[1]Sharvari Gaikwad, [2]Ruchi Singh, [3]Dr. Sidhharth Hariharan

[1]Bachelor of Engineering, [2] Bachelor of Engineering, [3]Associate Professor

[1]Dept. of Computer Engineering,

[1]Terna Engineering College, Nerul, Navi Mumbai, Maharashtra, India

Abstract: Text summarization, the process of condensing large volumes of text into concise, coherent, and informative summaries, holds immense significance in the digital age. This paper explores various extractive text summarization approaches that are the fusion of Natural Language Processing (NLP) and Machine Learning (ML) techniques to advance the state of the art in text summarization. Our system leverages models, to understand the context and semantics of the text. Through techniques like tokenization, attention mechanisms, and pre-trained language models, we extract key information from documents, ensuring that the generated summaries retain their original meaning and coherence. Machine Learning plays a pivotal role in our system by training on a vast corpus of human-generated summaries, allowing our model to learn the art of summarization through data-driven patterns. We delve into the intricacies of extractive and abstractive summarization methods, showcasing the benefits and challenges associated with each. We evaluate our system on diverse datasets, encompassing news articles, academic papers, and legal documents. The results demonstrate the efficacy of our approach in generating informative and coherent summaries across various domains, surpassing baseline methods.

Keywords **-** Extractive Summarization, Text Summarization, Summarization Methods, Comparison Study, Evaluation Metrics, and Text Processing**.**

Introduction

In an era of abundant information, the overwhelming volume of textual data from diverse sources presents a challenge in efficient information retrieval and knowledge extraction. Individuals and organizations are inundated with vast amounts of text, including news articles, research papers, legal documents, and business reports. This huge and overwhelming volume hampers effective decision-making, knowledge management, and content curation.

The data produced cannot be processed readily as it comes, this is because there are anomalies present in the data that can hinder the results. One potential tool for addressing the large data issue is text summarization. Instead of forcing analysts to personally search through long text themselves for information, we can apply machine learning algorithms to the documents within a corpus and reduce them to only a few representative sentences [8]. This paper describes various methods used for extracting information from a long text into a concise summary. The data is cleaned and pre-processed to have correct and complete data that is modeled against different extractive text summarisation approaches. Extractive text summarization is divided into two phases: 1) Pre-processing and 2) Processing [11].

These models are evaluated based on the ROGUE-n scores to help find the ideal results.

I. Related Work

In this section, we discuss the different approaches and how they come to be in the field of computerized synopsis [2]. Sentence Ranking method and Text Rank algorithm are proposed and compared against each other, where the end output is converted into audio. It is seen that Text Rank performs better compared to the sentence ranking method [1]. Another system that uses the concept of sentence ranking, the paper pre-processes the textual input through tokenization and removes stop words, the tokenized words are weighted to get the sum of weighted frequencies. The sentences from the high-weighted frequencies are then converted into audio form. The results are compared against MS Word-generated and human-generated summaries [11].

A corpus of 3 datasets, CNN Dataset, Blog Summarization dataset, and SUMMAC dataset are pre-processed using 15 different sentence scoring techniques, these techniques are broadly classified into word scoring, sentence scoring, and graph scoring. All 15 techniques are implemented and evaluated based on quantitative and qualitative evaluation [16].

TextRank, a graph-based ranking model for text processing. It proposes unsupervised methods for keyword and sentence extraction, demonstrating favorable results compared to previous benchmarks. The model applies graph-based ranking algorithms to lexical or semantic graphs extracted from natural language documents, allowing for various natural language processing applications [17]. Automated text summarization tools for complex text data like medical records, scientific papers, and legal documents. It compares three different summarization approaches - TF-IDF, TextRank, and LDA - for different input domains based on their complexity. The evaluation of the generated summaries is done using ROUGE metrics [7].

Semantic Analysis-based Approaches and Graph-based Approaches are utilized to produce concise and relevant summaries from single or multiple documents. Techniques include extractive and abstractive summarization methods. Evaluation methods play a crucial role in assessing the quality of automatic summaries [13].

The SpaCy library creates word vectors that use the principle of object-oriented programming. This converts the text into an object as a whole, which creates word vectors. Word vectors help in the proper assignment of real numbers to represent the meaning and efficiency of the words and also to cluster them based on Mathematical operations [6].

Compression and the Retention Ratios that normally extractive summarizers get wrong or disregard and thus the resulting summary loses important information. The proposed system first tokenizes the sentences, removes stop words, generates a bag of words mode, computes the mean of the sentences, computes standard deviation, and removes outliers by computing the z-score. The generated summary contains extracted facts and sentences based on keywords. This summary is compared with the original document text and shows it performed better than the other extractive summarizers and the LDA model while keeping in mind compression and retention ratio in mind respectively [3].

II. Approaches

Word Frequency Method

The word frequency method is used to get the summary of the text by computing the weighted frequency for each word used in the content. With the help of weighted frequency, each sentence is assigned a score, and a threshold value is computed. By changing the threshold value of the sentence score, a different summary can be obtained. While simple and computationally efficient, it serves as a useful baseline for extractive summarization tasks and can be augmented with other techniques for improved accuracy [4]. SpaCy python library is used to implement the word frequency algorithm.

Sentence Frequency Method

The sentence frequency method ranks sentences based on their occurrence frequency. After tokenizing the text into sentences, each sentence's frequency is calculated. Higher-frequency sentences are prioritized, assuming their importance. This approach selects the top-scoring sentences to form the summary, condensing the text while emphasizing recurring information. However, it may overlook semantic nuances and context, potentially resulting in less coherent summaries.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a summarization technique that assesses word importance in a document collection. The frequency of each word (TF) in a document is calculated and its uniqueness across the collection (IDF). Multiplying these values gives the TF-IDF score, highlighting words significant to a specific document but rare across the collection. In summary, sentences are scored based on the aggregated TF-IDF scores of their words, selecting those with the highest scores for inclusion in the summary. The priority is those sentences containing key terms unique to the document, enhancing the summary's relevance and informativeness.

TextRank

TextRank, an unsupervised graph-based ranking algorithm, operates by representing the text as a graph, with sentences as nodes and edges weighted based on similarity metrics like cosine or Jaccard similarity. Employing the PageRank algorithm, akin to Google's web page ranking, TextRank assigns importance scores to each sentence iteratively. Higher-ranking sentences are deemed more vital and are chosen for inclusion in the summary. This approach excels in identifying significant sentences by considering their relationships with others in the text.

Latent Semantic Analysis (LSA)

LSA is utilized to uncover underlying structures and extract key concepts from document collections. The text is represented in a high-dimensional vector space, followed by dimensionality reduction through singular value decomposition (SVD). This process identifies latent semantic concepts that capture relationships between terms and documents. Sentences are then scored based on their association with these concepts, with higher-scoring sentences selected for the summary..

III. IMPLEMENTATION

In the system, the data is meticulously pre-processed to ensure data cleanliness and uniformity. Different extractive summarizer approaches are implemented that select sentences based on relevance and sentence scores.

The data is pre-processed. In this phase, we tokenize, lowercase, and remove stop words. Rouge scores are used as a performance metric to evaluate the performance of all methods. The Rouge Score is a set of criteria used to assess the quality of artificial text summarising systems. The results of the evaluation metrics will be compared for each different technique and model. Following evaluation, the algorithms constructed are delivered to the web application and allow any user to generate summaries from textual input. The web application asks the user for input text which can either be a block of text or a text document. Figure 1 shows the flow diagram for the web application..

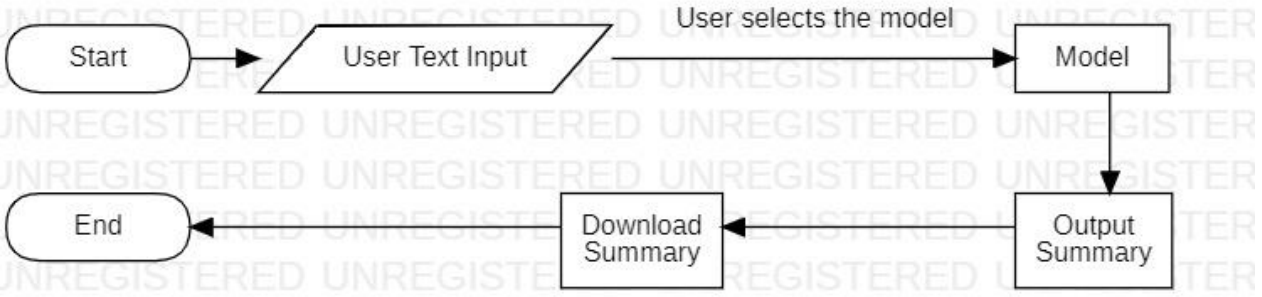

Fig 1: Flow diagram of application

IV. RESULTS AND DISCUSSION

The integrated approach for summary generation is implemented in Google Colab. The Web application is made with Streamlit where the original review text is given as input and a generated summary with the LSA Algorithm as shown in Figure 2.

Original Text: *Modern-day cats descended from a subspecies of African wildcat, Felis silvestris lybica, which today is the most common and widespread wildcat. Thousands of years ago, these wildcats were likely drawn to human settlements and their plentiful mice and food scraps.*

Summary: *Modern-day cats descended from a subspecies of African wildcat, Felis silvestris lybica, which today is the most common and widespread wildcat.*

Fig 2: Example of sample text and its summary

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are widely used to evaluate the quality of generated summaries. The performance metrics used are ROUGE-1, which measures unigram overlap, assessing precision, recall, and F1 score based on shared individual words, ROUGE-2 extends this to bigram overlap, capturing pairs of consecutive words. Rouge-N counts how many n-grams in the system produced text and our summaries match each other. Bigram is made up of successive words, whereas a unigram consists of just one [1]. ROUGE-L emphasizes the longest common subsequence of words, evaluating summary quality based on precision, recall, and F1 score relative to reference summaries. ROUGE-LSUM, a variant of ROUGE-L, operates at the summary level, considering the length of both generated and reference summaries. All the methods are evaluated against Rogue scores and the results are displayed in the Table 1, and a graph Figure 3 is drawn to demonstrate the results.

Table 4.1: Results of evaluation

Methods	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
Word Frequency	0.58	0.56	0.29	0.53
Sentence Frequency	0.58	0.56	0.58	0.56
TextRank	0.41	0.4	0.3	0.41
TF-IDF	0.6	0.57	0.24	0.24
LSA	0.58	0.55	0.58	0.58

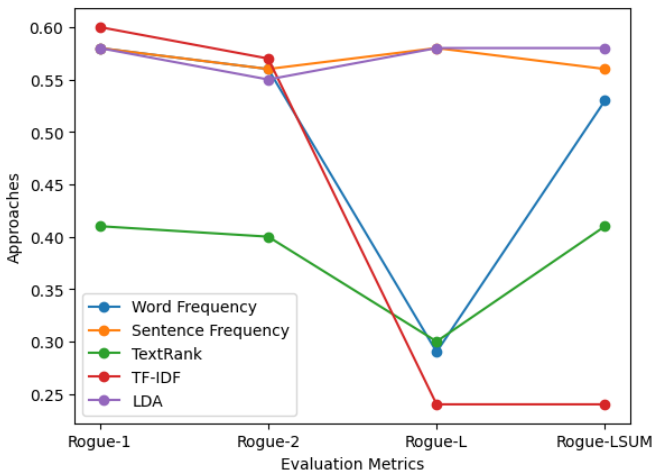

Fig 3: Line graph for approaches against evaluation metrics

V. CONCLUSION

The comparative study of different extractive text summarization techniques underscores the diverse approaches available for condensing text while retaining its key information. Through our examination, we observed the efficacy of methods such as frequency-based summarization, TextRank, and Latent Semantic Analysis (LSA) in capturing salient content.

While frequency-based methods offer simplicity and ease of implementation, graph-based techniques like TextRank leverage semantic relationships to produce coherent summaries. Moreover, LSA, through dimensionality reduction and concept identification, facilitates a deeper understanding of text structures, leading to more informative summaries. By comprehensively analyzing these techniques, we can better appreciate their strengths and limitations, paving the way for further advancements in extractive summarization research and applications.

REFERENCES

[1] M. Majeed and K. M. T, "Comparative Study on Extractive Summarization Using Sentence Ranking Algorithm and Text Ranking Algorithm," 2023 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, India, 2023, pp. 1-5, doi: 10.1109/PICC57976.2023.10142314

[2] Gulati, V.; Kumar, D.; Popescu, D.E.; Hemanth, J.D. Extractive Article Summarization Using Integrated TextRank and BM25+ Algorithm. Electronics 2023, 12, 372. https://doi.org/10.3390/ electronics12020372

[3] Waseemullah; Fatima, Z.; Zardari, S.; Fahim, M.; Andleeb Siddiqui, M.; Ibrahim, A.A.A.; Nisar, K.; Naz, L.F. A Novel Approach for Semantic Extractive Text Summarization. Appl. Sci. 2022, 12, 4479. https://doi.org/10.3390/ app12094479

[4] AbinayaN, AnbukkarasiS and VaradhaganapathyS "Extractive Text Summarization Using Word Frequency Algorithm for English Text", Information Retrieval Evaluation, December 9-13, 2022, India

[5] S. JUGRAN, A. KUMAR, B. S. TYAGI and V. ANAND, "Extractive Automatic Text Summarization using SpaCy in Python & NLP," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 582-585, doi: 10.1109/ICACITE51222.2021.9404712.

[6] Rani, U., & Bidhan, K. (2021). Comparative Assessment of Extractive Summarization: TextRank, TF-IDF and LDA. Journal of scientific research.

[7] Shearing, S., Gertner, A., Wellner, B., & Merkhofer, L. (2020). Automated Text Summarization: A Review and Recommendations.

[8] M. S M, R. M P, A. R E and E. S. G SR, "Text Summarization Using Text Frequency Ranking Sentence Prediction," 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2020, pp. 1-5, doi: 10.1109/ICCCSP49186.2020.9315203.

[9] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[10] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040

[11] Khan, Rahim & Qian, Yurong & Naeem, Sajid. (2019). Extractive based Text Summarization Using KMeans and TF-IDF. International Journal of Information Engineering and Electronic Business. 11. 33-44. 10.5815/ijieeb.2019.03.05.

[12] Pradeepika Verma, Sukomal Pal, and Hari Om. 2019. A Comparative Analysis on Hindi and English Extractive Text Summarization. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 18, 3, Article 30 (September 2019), 39 pages. https://doi.org/10.1145/3308754

[13] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2017, pp. 1-6, doi: 10.1109/ICCCSP.2017.7944061

[14] Ed Collins, Isabelle Augenstein and Sebastian Riedel "A Supervised Approach to Extractive Summarisation of Scientific Papers", 2017, Journal CoRR, url{http://arxiv.org/abs/1706.03946}

[15] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, Luciano Favaro,Assessing sentence scoring techniques for extractive text summarization, Expert Systems with Applications, Volume 40, Issue 14, 2013,Pages 5755-5764, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2013.04.023.

[16] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

[17] https://cran.r-project.org/web/packages/textrank/vignettes/textrank.html

[18] https://blog.peiyingchi.com/2019/10/29/TextRank-LexRank-DivRank/