Analysis of Extractive Summarization Methodologies for Single Document Texts

Kajal Agrawal[1] Dr. Sharvari Tamane[2]
[1]*Associate Professor, MGM University Nath School of Business and Technology, Aurangabad (MS), India.*
[2] *Professor and HOD- MGM Department of Information & Communication Technology.*

Abstract:
In the big data era, text summarizing is essential for reducing long documents into short, easily understood summaries and promoting effective information consumption. An overview of current trends, varieties, and difficulties in text summarizing is provided in this study. A thorough literature review is followed by an exploration of the field's many methodologies and approaches, which include models that incorporate both local and global context, hybrid techniques that combine unsupervised and fuzzy logic, and supervised methods that use neural networks. Along with domain-specific, query-based, and generic summary strategies, the study looks at how extraction-based and abstractive summarization functions. It also explores the workings of text summarization algorithms, including Text Rank and Sequence-to-Sequence Modeling, highlighting how well they can summarize textual content. The significance of swarm intelligence in improving text summarization techniques is also covered in the paper. Lastly, it emphasizes the benefits of text summarization, such as time savings, multilingual compatibility, and increased productivity in understanding and information retrieval. This study will help to enhance our knowledge and comprehension of text summarizing techniques and the range of effective uses they may have for managing enormous volumes of textual data.

Keywords: Extractive Summarization, Single Document, Methodologies, Comparative Analysis, Evaluation, Swarm Intelligence

I) Introduction:

Introduction to Text Summarization

The process of condensing lengthy publications into digestible paragraphs or sentences is known as text summarization. The process ensures that the paragraph's meaning is maintained while simultaneously extracting pertinent information. This reduces the amount of time needed to understand lengthy resources, such as research articles, without sacrificing important details.

Text summarizing is the act of creating a succinct, coherent, and fluid summary of a longer text document while emphasizing the key elements of the text. Text summarizing is the act of creating a succinct, coherent, and fluid summary of a longer text document while emphasizing the key elements of the text. Text identification, interpretation, summary production, and examination of the generated summary are some of the challenges associated with text summarization. In extraction-based summarizing, one of the most significant tasks is identifying key terms in the document and using them to unearth pertinent information to include in the summary.

Requirement of text summarization

In the age of big data, the volume of text data that is accessible from multiple sources has skyrocketed. There is a lot of knowledge and experience in this lengthy text, which needs to be sufficiently condensed in order to be helpful. A great deal of research in natural language processing (NLP) is needed for automatic text summarization because the number of documents available is increasing. Automatic text summarizing is the task of preserving the meaning of the original text content while producing a concise and fluid summary without human aid. It's challenging because, in order to create a summary that highlights the key aspects of a piece of literature, we often read it through to gain a deeper understanding of it. Automated text summarization is a difficult and time-consuming process since computers cannot understand human language or comprehension. Moreover, text summarization expedites research, reduces reading time, and increases the amount of information that may fit in a given area. Text summary also speeds up research, cuts down on reading time, and expands the quantity of information that may be included in a given space.

II) Objectives of the Study:

1.To present contemporary trends and the principles of swarm intelligence optimization algorithms (PSO, ACO) and their applicability to text summarization.
2. To explore different types of Extractive Text Summarization and its advantages.
3. To underscore on the nuances of Inverse Document Frequency (IDF)
4. To lay emphasis on Text Rank Algorithm
5. Evaluate the performance of enhanced text summarization techniques using real-world datasets and standard evaluation metrics (e.g., ROUGE).

III) Research Methodology

This study is curated based on the secondary sources for its data, it emphasizes an investigative approach to its subject. Electronic journals, websites, and textbooks are the sources of the data, which helps to provide a thorough and in-depth analysis of the selected subject.

IV) Literature Review:

(Salima et al, 2019) [1] Conducted series of systematic experiments in their research study, “Supervised Method for Extractive Single Document Summarization” based on Sentence Embedding’s and Neural Networks to compute sentences scores, they used a feed forward neural network (FFNN) that exploits both sentence vector representations and the centroid embedding’s vector of the document and additional features that capture relations between them. After scoring each sentence of the input document, they generated its summary by selecting and concatenating the top-ranked sentences. It highlighted the intrinsic differences of short and long document datasets and showed that summarizing long documents requires extra compression of the source text through the identification of key narratives that are more uniformly scattered across the source documents.

(Som Gupta and S.K Gupta, 2018) [2] In their research study **“**A Hybrid Approach to Single Document Extractive Summarization”, presented a hybrid approach to automatic text summarization that makes use of fuzzy logic, feature-based extraction, semi-supervised technique MMR, and unsupervised approach PageRank. While fuzzy logic helps handle summaries' uncertainty, feature-based extraction aids in capturing word-level relationships, and PageRank aids in capturing sentence-level associations. According to the experimental results achieved using the ROUGE framework, the combined method outperforms the individual approaches in terms of quality or comparable results. Even though the hybrid strategy produces superior outcomes, the precision can occasionally be as low as 1%.

(Wen Xiao and Giuseppe Carenini, 2019) [3] While examining the “Extractive Summarization of Long Documents by Combining Global and Local Context” , proposed their models based on datasets that found a notable improvement in performance when local context, or topic information, is added. In cases of lengthy documents, the improvement is even more pronounced. Performance is never appreciably enhanced by include a global context representation of the entire content. Essentially, it appears that all of our model's advantages even for the longest documents come only from simulating the local context. Additionally, it demonstrates that news articles tend to be less biased than scientific publications; that is to say, an extractive summary should not be formed from the first few phrases of these pieces. If we take the first five sentences from the conclusions section if we set the summary length restriction to the length of our abstract. Two additional sentences are recovered if the length is increased to 200 words, and they do appear to offer helpful supplementary information.

(Anusha Pai , 2014) [4] presented a very insightful article on “Text Summarizer Using Abstractive and Extractive Method” , where in there was assessment of her model on two sizable scientific article datasets, which comprise documents that are significantly lengthier than in previously utilized corpora, and compare it with prior studies in both extractive and abstractive summarization.

(Martha & et al, 2015) [5] In their research article "Extractive Single-Document Summarization Based on Global-Best Harmony Search and a Greedy Local Optimizer", presented ESDS-GHS-GLO, a novel mimetic algorithm based on greedy search and global-best harmony search that generates extractive summaries automatically from a single page. In this problem, the agent is represented by a large number of "zeroes" and a small number of "ones" (phrases chosen for the summary); alternatively, the agent can be implemented as a list that only contains the chosen sentences. The Global-best Harmony Search algorithm simplifies the algorithm design process by eliminating the need for the selection, crossover, mutation, and replacement processes that are typical in genetic algorithms.

(Nazreena & Bhogeswar, 2021) [6] Derived out of their research work "Query-Based Extractive Text Summarization Using Sense-Oriented Semantic Relatedness Measure", and proposed a redundancy free query based extractive text summarization method. Under unsupervised learning methods, the proposed method provides excellent extraction ability and better query based summary quality even compared with some supervised methods. They also emphasized on query relevance performance with other query-based text summarization system and have also introduced a word sense disambiguation method for query-based text summarization. This method helps in finding the sense-oriented query relevance sentences on the basis of its meaning. The evaluation process has an impact on the query-based text summarization method for numerous text documents overall. The summary produced by DUC datasets is limited to 250 words. One significant flaw in the text summarization evaluation system is that it is quite challenging to produce a summary in just 250 words that is consistent with a summary created by a human.

(Yang, Ivan & Mirella Lapata, 2019) [7] In their research paper "Single Document Summarization as Tree Induction", concluded that in addition to classifying phrases as summary-worthy or not, their summarizer which referred to as SUMO, short for Structured Summarization Model also induces the structure of the source material as a multi-root tree. SUMO, performs better than variations with document attention as well as a basic Transformer model without any document attention. Overall, SUMO with three levels of structured attention outperforms other models, supporting our theory that document-level structure aids in summarizing. Findings also show that LEAD-3 performs worse across the board than SUMO and all Transformer-based models with document attention (doc-att). SUMO (3-layer) approaches are on par with or superior to cutting-edge methods.

According to **(ZhaoningLi et al. 2009),** [8] causality extraction is approached as a sequence tagging problem in their paper, with a proposed solution using self-attentive BiLSTM-CRF. Specifically, they introduce SCITE as a method for extracting causality from natural language text, based on their causality tagging scheme. To address the challenge of insufficient data, they adopt a strategy of transferring Flair embeddings trained on a large corpus to their task.

(Ramya and Kiran 2022) [9] In their research based on "Privacy Preserving Text Document Summarization" asserted that Extensive domain-specific text pre-processing is required prior to the privacy preserved summary generation. The results indicate that the proposed tf-idf-based summarization computationally performed well when compared with other summarization techniques. It also preserves patient privacy without defying privacy constraints.

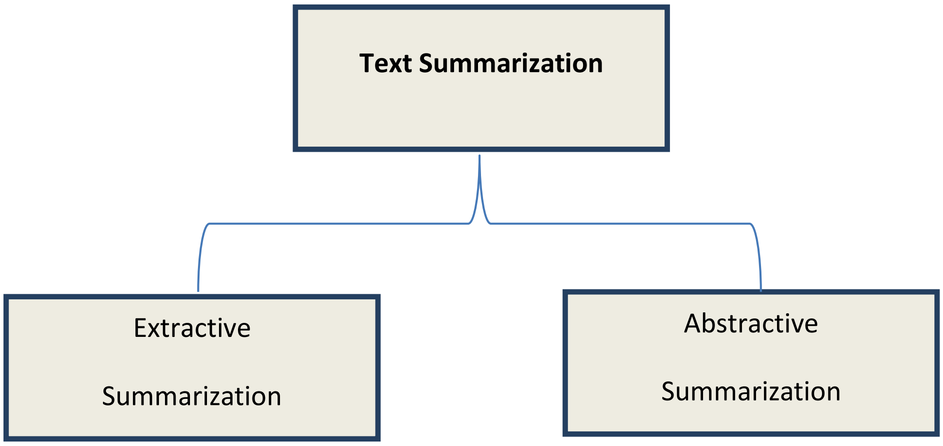
(Ke chin & et el 2019) [10] In conclusion, this research introduces HPSO-SSM, a novel wrapper-based approach for feature selection within an unsupervised framework utilizing a modified PSO algorithm. Addressing key limitations of prior PSO-based methods, such as insufficient algorithm diversity and a lack of balance between exploration and exploitation, HPSO-SSM integrates innovative adjustments including a logistic map sequence for inertial weight tuning and a spiral-shaped mechanism for enhanced search quality.

In a research on " News keywords extraction algorithm based on TextRank and classified TF-IDF" **(Ao et al, 2020)** [11] pinpointed that FSL-TR introduces the LSTM classification model, it proves to be less time-consuming compared to TF-IDF and TextRank algorithms, particularly when applied to news texts and requiring a higher recall rate within seconds. Despite the inherent complexity in keyword extraction, the TFSL-TR algorithm surpasses the traditional approaches, showcasing its effectiveness in enhancing the accuracy of keyword extraction. While TextRank remains a widely used algorithm in this field, the TFSL-TR presents a significant advancement towards achieving superior keyword extraction accuracy, marking a promising direction for further research and development in text processing algorithms.

(Madhuri & Kumar 2019) [12] In their research article "Extractive Text Summarization Using Sentence Ranking" address the complexity of automatic text summarization, highlighting its various sub-tasks, each with the potential to yield high-quality summaries. Emphasizing the significance of identifying essential paragraphs within a given document in extractive summarization, they propose a novel statistical approach for sentence ranking. This method selects sentences based on their rank to generate a summarized text, subsequently converting it into audio format. The authors assert that their proposed model demonstrates enhanced accuracy compared to traditional approaches.

V) Different types of Text summarization

There are mainly two types of text summarization in NLP:



(Fig: 1.1- Types text summarization)

Extraction based summarization

The process of extractive text summarizing involves taking the most important words out of a source text and putting them together in a summary.

The extraction is carried out in accordance with the specified measure without causing any changes to the texts.

This method finds important passages in the text, cuts them out, and then sews the text back together to produce a condensed version.

Abstractive Summarization

Abstractive summarization is an additional method of text summarizing. In this phase, we take the original content and turn it into new sentences. Unlike this, we only used the terms that were present in our earlier extraction approach. It's probable that the original text does not contain the phrases created by abstractive summarization. The grammatical mistakes of the extractive method can be avoided when abstraction is utilized for text summarization in deep learning problems. Compared to extraction, abstraction is more efficient. However, the text summarizing algorithms required for abstraction are more difficult to construct, which is why extraction is still commonly utilized.

On the basis of Context

Domain-Specific

Domain expertise is used in domain-specific summarization. Domain-specific summarizers can be used to combine contextual information, language, and specific expertise. For instance, models can be used in conjunction with medical terminology to improve comprehension and summarization of scientific publications.

Query-based

Natural language queries are the main focus of query-based summaries. This is comparable to the Google search results. Sometimes, when we enter queries into Google's search bar, it brings us webpages or articles that address our queries. It presents an excerpt or synopsis of an article pertinent to the search term we typed in.

Generic

In contrast to domain-specific or query-based summarizers, generic summarizers are not preprogrammed to make any assumptions. It is merely a summary or condensed version of the original text.

VI) Working of text summarization algorithm

In natural language processing, text summarization is usually treated as a supervised machine learning problem. Develop a technique for removing the crucial keys from the source document. Gather text documents that have well-labeled keywords. The keys and the designated extraction technique must work together. To increase accuracy, negatively labelled keys can also be constructed. Use a binary machine learning classifier to train in order to generate the text summary. Lastly, create every pertinent word or phrase in the test phrase and categorizes it appropriately.

Sequence-to-Sequence Modeling (Seq2Seq)

For the solution of any sequential data problem, we can employ a Seq2Seq model. Popular sequential information applications include named entity recognition, neural machine translation, and sentiment categorization. When using neural machine translation, the input is a text in one language, and the output is likewise a text in another language. A list of words is the input for Named Entity Recognition, and the output is a list of tags for every word in the list. The encoder and decoder are the two main parts of the Seq2Seq modeling process. Let's examine this idea of these:

Encoder;

One word is sent into the encoder at each time step, and the encoder—a Long Short Term Memory model (LSTM—reads the entire input sequence. After that, the data is analyzed at each time step, and the contextual data from the input sequence is recorded.

Decoder

Similar to the decoder, it is an LSTM network that predicts a sequence that is one time step delayed after analyzing the entire target sequence word-by-word. The decoder is taught to predict the following word in the sequence given the preceding word.

Requirement of text summarization

In the big data age, the volume of text data that is accessible from multiple sources has skyrocketed. There is a lot of knowledge and experience in this lengthy text, which needs to be sufficiently condensed in order to be helpful. A great deal of research in natural language processing (NLP) is needed for automatic text summarization because the number of documents available is increasing. Automatic text summarizing is the task of preserving the meaning of the original text content while producing a concise and fluid summary without human aid. It's challenging because, in order to create a summary that highlights the key aspects of a piece of literature, we often read it through to gain a deeper understanding of it.

Automated text summarization is a difficult and time-consuming process since computers cannot understand human language or comprehension. Moreover, text summarization expedites research, reduces reading time, and increases the amount of information that may fit in a given area.

VII) Inverse Document Frequency (IDF)

S Kavita Ganesan has elaborately discussed the concept of Document Frequency in which she asserts that Inverse Document Frequency (IDF) indicates how frequently a word is used. Its score decreases with increased usage across papers. The term loses significance as its score drops. For instance, the word "the," which is found in practically all English writings, has a relatively low IDF score since it contains very little information on "topics." By comparison, although the term "coffee" is commonly used, its usage is not as widespread as that of the word "the." Coffee would therefore score higher on the IDF than the. IDF is typically calculated as follows:

Illustration: If a word occurs 10 times in a document and its IDF weight is 0.1, the document's score would be 1 (10*0.1=1). Now, the score would be 5 if the word "coffee" also occurred 10 times and had an IDF weight of 0.5. Coffee would come before the word the in the ranking of the terms based on the scores (in descending order, of course!), suggesting that coffee is more significant than the word the. IDF is a handy little formula that may be used for a variety of tasks, including keyword extraction, feature weighting in text classifiers, and stop-word list creation.

VIII) Text Rank Algorithm

The Text Rank algorithm is a graphical technique that divides pre-processed data into sentences and words, which serve as graph vertices. The similarity between the phrases and words determines the weight of the edges connecting them. A similarity matrix is created in order to calculate this similarity Text Rank's main benefit is that it is an unsupervised graph-based algorithm, meaning it can determine the key elements of a textual content without the need for a human summary or training dataset. Compared to comparable supervised algorithms, unsupervised algorithms require less manual data preparation, which saves time overall. Text Rank is an ATS system that operates on word occurrence and is language-independent, in addition to being an unsupervised algorithm. It determines the significance of words and sentences and only includes the most illuminating statements in the output summary.

PageRank Algorithm

We would need to calculate a number known as the PageRank score in order to rank these pages. The likelihood of a user viewing that page is represented by this score.

The following stages illustrate how the probability were initialized:

1. The starting value of the probability of moving from page i to page j, or M[i][j], is 1/(number of unique links in web page wi).

2. The likelihood will start at 0 if there is no connection between pages I and J. 3.It is considered that a user has an equal chance of transitioning to any page if he lands on a dangling page. The initialization of M[i][j] will therefore be 1/(number of web pages).

3. The rationale behind Text Rank is similar to that of PageRank, however modified slightly: Web Pages are replaced with text sentences. Rather than 1/total_links from Page B to A, the similarity matrix for index [A, B] is populated with similarity values between sentences A & B. 4. The cosine similarity between two sentences or the number of most common terms between two sentences could determine the similarity values. Word embedding provides the vector representation of words, whereas cosine similarity refers to the similarity between two vectors as evaluated by the cosine of the angle between two vectors and assesses if two vectors are pointing in roughly the same direction.

Dataset Description

The CSV file that serves as the dataset for this challenge has four distinct sources properties, which are specified as follows:

1. article id;
2. article title;
3. text;
4. source of article

However, the third attribute—the article text—will be the focal point of our issue. To create a single summary, we will take into account the article text from each row.

Dataset Preprocessing

1. Fill a Pandas data frame with CSV data.
2. Put them all into a list of sentences and take into account the article text attribute.

3. Eliminate special characters, digits, and punctuation.
4. Lowercase alphabets.
5. Remove any stop words that are in the sentences. Eventually, with the aid of the Glove word vectors, we will have a list of clear sentences from which to construct vectors for sentences in our data.

Implementation of the Text Rank Algorithm

Actions to take following the preprocessing stage:

1. Using word embedding to represent sentences as vectors.
2. Creating a matrix of similarities.
3. Graph/network conversion of the Similarity Matrix for PageRank Algorithm use.
4. The PageRank Algorithm's application.
5. Using the top N sentences and their rankings to extract a summary.

Swarm intelligence (SI) is the collective intelligence that arises from the collective behaviour of simple people interacting with their surroundings and with each other locally, leading to the emergence of coherent, functioning global patterns . The two main computational components of swarm intelligence are Particle Swarm optimization (PSO), which is inspired by the social behaviour of fish schools or flocks of birds, and Ant Colony Optimization (ACO), which is inspired by ant behaviour.

IX) Advantages of Text Summarization

Text summarization offers a multitude of benefits across various domains. Firstly, it enables time-saving by allowing readers to quickly grasp the main points of a document without the need to read through the entire text. This is particularly beneficial in scenarios where individuals are pressed for time or need to efficiently process a large volume of information. Additionally, summarization tools enhance efficiency in information retrieval by condensing lengthy documents into concise summaries, thereby aiding in navigating through the era of information overload. These summaries not only improve comprehension by providing a clear overview of the main ideas and arguments but also facilitate cross-language communication due to the language-independent nature of text summarization techniques. Moreover, with advancements in natural language processing and machine learning, automated summarization processes have become scalable, customizable, and invaluable for decision support and content creation purposes. Summaries serve as decision-making aids by offering concise information essential for informed decisions and can also be repurposed to create various types of content, enhancing their versatility and utility in diverse contexts.

X) Conclusion

Within the field of big data, text summarizing is a crucial tool for condensing large volumes of textual data into easily readable summaries. With an emphasis on both extraction-based and abstractive methods, this paper offers a thorough analysis of text summarizing strategies spanning from conventional statistical methods to state-of-the-art deep learning models. The paper highlights techniques like supervised neural networks and cutting-edge algorithms like Text Rank, while illuminating the intrinsic difficulties in text summarization, such as preserving coherence and addressing redundancy, through a thorough literature survey and comparative analysis. It also emphasizes the importance of swarm intelligence and the critical role that context plays in optimizing summarization processes through Particle Swarm Optimization and Ant Colony Optimization, regardless of whether the context is domain-specific or query-based. The study establishes a strong basis for future developments in the subject by clarifying current trends and investigating different methodological approaches. Text summarization becomes an indispensable instrument for effective information intake, increased productivity, and knowledge acquisition in the face of the difficulties presented by the big data era.

References:

1. *Salima Lamsiyah, Abdelkader El Mahdaouy, Said El Alaoui Ouatik, Bernard Espinasse (2019) A Supervised Method for Extractive Single Document Summarization based on Sentence Embeddings and Neural Networks*

,International Conference on Advanced Intelligent Systems.

2. Som Gupta et al, (2018) International Journal of Computer Science and Mobile Computing, Vol.7 Issue.11, November- 2018, pg. 142-149 ISSN 2320–088X IJCSMC, Vol. 7, Issue. 11, November 2018, pg.142 – 149

3. Wen Xiao and Giuseppe Carenini (2019), Extractive Summarization of Long Documents by Combining Global and Local Context, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

4. Anusha Pai (2014), Text Summarizer Using Abstractive and Extractive Method International Journal of Engineering Research & Technology (IJERT)- ISSN: 2278-0181- Vol. 3 Issue 5, May - 2014

5. Martha Mendoza , Carlos Cobos, and Elizabeth León (2015) Extractive Single-Document Summarization Based on Global-Best Harmony Search and a Greedy Local Optimizer, Springer International Publishing Switzerland 2015 O. Pichardo Lagunas et al. (Eds.): MICAI 2015, Part II, LNAI 9414, pp. 1–15, 2015.

6. Rahman, Nazreena & Borah, Bhogeswar. (2021). Query-Based Extractive Text Summarization Using Sense-Oriented Semantic Relatedness Measure. 10.21203/rs.3.rs-1102477/v1.

7. Yang Liu, Ivan Titov and Mirella Lapata (2019), Single Document Summarization as Tree Induction,Proceedings of NAACL-HLT 2019, pages 1745–1755 Minneapolis, Minnesota, June 2 - June 7, 2019. c 2019 Association for Computational Linguistics

8. Li, Z., Li, Q., Zou, X., & Ren, J. (2019). Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings. ArXiv, abs/1904.07629.8491.

9. Shree, A N & P, Kiran. (2022). Privacy Preserving Text Document Summarization. Journal of Engineering Research and Sciences. 1. 7-14. 10.55708/js0107002.

10. Chen, Ke & Zhou, Feng-Yu & Yuan, Xian-Feng. (2019). Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. Expert Systems with Applications. 128. 140-156. 10.1016/j.eswa.2019.03.039.

11. Ao, Xiong & Yu, Xin & Liu, Derong & Tian, Hongkang. (2020). *News keywords extraction algorithm based on TextRank and classified TF-IDF. 1364-1369. 10.1109/IWCMC48107.2020.914*

12 Madhuri, J.N. & Kumar, R.. (2019). Extractive Text Summarization Using Sentence Ranking. 1-3. .1109/IconDSC.2019.8817040.

Web sources

1. kavita-ganesan.com
2. projectsgoal.com
3. ir.canterbury.ac.nz
4. cloudfront.escholarship.org
5.freepatentsonline.com
6 mdpi.com
7.ncbi.nlm.nih.gov