International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

Comprehensive Survey of Abstractive Text Summarization Techniques

Kundan Chaudhari[1], Raj Mahale[2], Fardeen Khan[3], Shradha Gaikwad[4], Kavita Jadhav[5]

[1-5]D Batu, Nutan College of Engineering and Research Vishnupuri Talegaon Dabhade, India.

Email ID: *kundanchaudhari8329@gmail.com[1], rajmahale04@gmail.com[2], khanfardeen212@gmail.com[3], shradhagaikwad3752@gmail.com[4], kavitajadhav@ncer.in[5]*

Abstract

Text summarization using pre-trained encoders has become a crucial technique for efficiently managing large volumes of text data. The rise of automatic summarization systems addresses the need to process ever-increasing data while meeting user-specific requirements. Recent scientific research highlights significant advancements in abstractive summarization, with a particular focus on neural network-based methods. A detailed review of various neural network models for abstractive summarization identifies five key components essential to their design: encoder-decoder architecture, mechanisms, training strategies and optimization algorithms, dataset selection, and evaluation metrics. Each of these elements is pivotal in enhancing the summarization process. This study aims to provide a thorough understanding of the latest developments in neural network-based abstractive summarization models, offering insights into the evolving field and underscoring the associated challenges. Qualitative analysis using a concept matrix reveals common design trends in contemporary neural abstractive summarization systems. Notably, BERT-based encoder-decoder models have emerged as leading innovations, representing the most recent progress in the field. Based on the insights from this review, the study recommends integrating pre-trained language models with neural network techniques to achieve optimal performance in abstractive summarization tasks. As the volume of online information continues to surge, the field of automatic text summarization has garnered significant attention within the Natural Language Processing (NLP) community. Spanning over five decades, researchers have approached this problem from diverse angles, exploring various domains and employing a multitude of paradigms. This survey aims to delve into some of the most pertinent methodologies, focusing on both single-document and multiple-document summarization techniques, with a particular emphasis on empirical methods and extractive approaches. Additionally, the survey explores promising strategies that target specific intricacies of the summarization task. Notably, considerable attention is dedicated to the automatic evaluation of summarization systems, recognizing its pivotal role in guiding future research endeavors.

Keywords: *Abstractive Text Summarization; Encoder; Interpreter; Drill; Courtesy.*

1. Introduction

The Text summarization plays a crucial role in simplifying the extraction of key information from extensive texts, allowing quick access to important details and addressing challenges related to summary evaluation criteria. With the ongoing development and success of automatic text summarization methods, which have demonstrated significant effectiveness across various languages, there is a growing need for thorough reviews and summaries of these techniques. In this comprehensive review, we explore the latest methodologies employed in text summarization, focusing on an in-depth analysis of the techniques used, the datasets leveraged, the evaluation metrics applied, and the challenges faced

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

by each approach. Our review delves into how each method addresses these challenges, highlighting the strategies implemented to overcome obstacles in the summarization process. This extensive overview aims to provide insights into the diverse methodologies utilized in recent text summarization techniques, offering a consolidated understanding of the advancements made, the tools and data employed, and the hurdles encountered in this domain. Understanding these aspects not only enhances our comprehension of the field but also paves the way for developing more effective and efficient text summarization methodologies. Numerous natural language processing (NLP) tasks, ranging from sentiment analysis to question answering, have benefited from pre-trained language models. Significant advancements in this domain include models like ELMo, GPT, and the more recent Bidirectional Encoder Representations from Transformers (BERT), which stand as state-of-the-art pre-trained models. Specifically, the BERT Transformer represents a comprehensive solution for encoding phrases and sentences. Its pre-training involves exposure to large volumes of text with unsupervised objectives such as masked language modeling and next-sentence prediction. Additionally, its flexibility allows for fine-tuning to suit various task-specific purposes. Abstract text summarization has numerous applications across various industries and domains. In academia and research, it helps researchers swiftly identify pertinent articles and papers, saving significant time and effort. In the legal sector, it assists lawyers and legal professionals in summarizing case law and legal documents, thus enhancing their efficiency. In the business world, it facilitates the summarization of reports, emails, and other business documents, enabling better decision-making and communication. Moreover, abstract text summarization is valuable in content curation and information retrieval, ensuring that users can access the most relevant information quickly. It also aids individuals with disabilities or limited literacy levels in accessing information more easily, thereby promoting inclusivity. In today's information-rich world, the ability to efficiently and accurately summarize large volumes of text is becoming increasingly essential. Abstract text summarization, a key area within natural language processing, seeks to fulfill this need by developing algorithms and techniques that can automatically produce concise and coherent summaries of text documents. These summaries distill the key information and meaning from the original text, enabling users to quickly extract relevant information without having to read the entire document. This research explores how pre-training a language model can enhance its ability to generate text summaries. Summarization, distinct from other NLP tasks, requires a deeper understanding of natural language beyond merely interpreting individual words or phrases. The primary objective remains to condense the original text while preserving its core meaning. There are two predominant paradigms in summarization: extractive and abstractive. Extractive summarization often functions as a binary classification task, determining whether a text segment (typically a sentence) should be included in the summary. Conversely, abstractive summarization involves language generation techniques, creating summaries that may contain new phrases and expressions not explicitly present in the source text. Our analysis highlights the significant progress made in neural network-based approaches to abstractive summarization. We examine the encoder-decoder architectures, training strategies, optimization algorithms, datasets, and evaluation metrics used in these models. Notably, BERT-based encoder-decoder models have emerged as leaders in the field, showcasing the most recent advancements. This comprehensive review advocates for integrating pre-trained language models with neural network techniques to achieve optimal performance in abstractive summarization tasks. By synthesizing insights from recent studies, we aim to offer a detailed understanding of the current landscape and the challenges that lie ahead, facilitating the development of more robust and sophisticated text summarization systems. One of the significant advancements in this domain includes the integration of pre-trained language models like ELMo, GPT, and BERT. These models have revolutionized natural language processing tasks, including text summarization, by providing a robust foundation for

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

encoding text. BERT, in particular, stands out due to its ability to understand context through bidirectional training, making it exceptionally powerful for generating coherent and meaningful summaries. Its pre-training on vast amounts of text using unsupervised objectives such as masked language modeling and next-sentence prediction allows it to capture intricate language patterns, which are crucial for abstract summarization tasks.

2. Literature Review

[1] Author: - Ani Nenkova and colleagues explore the application of clustering methods for summarization, particularly emphasizing the use of graph-based methods within each cluster. Their research involves an evaluation of 25 articles from five subfields of computational linguistics using the manual pyramid evaluation approach. The study underscores the importance of frequency in sentence selection and suggests that human-generated summaries are ften more effective than those produced automatically. The paper references various systems and studies related to document summarization, including techniques that leverage word similarities, TF-IDF weights, and log-likelihood ratios. It also highlights significant advancements in news and multi-document summarization. Nenkova's research emphasizes the critical role of clustering methods and the evaluation of summarization techniques, providing valuable insights into the complexities of summarizing scientific papers. The study points to potential advancements in news and multi-document summarization and highlights the significance of frequency in sentence selection. It also discusses the greater effectiveness of human summaries compared to automated ones. By referencing various systems and studies, the paper showcases the diverse approaches and challenges within the field of document summarization. [2] Author: - Ramesh Nallapati and colleagues delve into the realm of abstractive text summarization, focusing on novel models based on Attentional Encoder-Decoder Recurrent Neural Networks (RNNs). Their study begins by situating these models within the broader context of related research on abstractive text summarization, highlighting key differences between summarization and machine translation tasks. The authors present an off-the-shelf attentional encoder-decoder RNN, originally designed for machine translation, and demonstrate its superior performance on two distinct English datasets. The paper introduces several innovative models to address specific challenges in summarization, such as modeling key words, capturing sentence-to-word rankings, and generating unusual or unseen words, all of which lead to further performance improvements. Additionally, the research introduces a new dataset specifically for abstractive summarization and establishes benchmarks for future studies. Key contributions of this work include the application of the attentional encoder-decoder RNN to summarization tasks, the development of novel models to enhance performance, and the creation of a new dataset for the field. The paper is organized to provide a detailed description of each specific problem in abstractive summarization and the novel models proposed to address these issues. The authors also present the results of their tests on three different datasets, followed by a qualitative analysis of the models' final outputs and concluding remarks on future research directions in the field. [3] Author: - Shashi Narayan and colleagues introduce the concept of extreme summarization, a novel document summarization task requiring abstractive modeling methods. Unlike traditional summarization, extreme summarization aims to generate a concise, one-sentence news summary that answers the question, "What is the article about?" This task explicitly avoids extractive techniques. To support their research, the authors compile a large real-world dataset by extracting online articles published by the British Broadcasting Corporation (BBC), each accompanied by a first-sentence summary. In response to the unique challenges of extreme summarization, the authors propose a topic-conditioned neural model based solely on convolutional neural networks. This model is designed to capture long-term dependencies within documents and identify relevant content effectively. Through experimental evaluations, they demonstrate that their proposed architecture surpasses an oracle extractive system and state-of-the-art abstractive approaches in both automatic and human assessments. The study's findings underscore the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

importance of high-level document understanding, especially regarding topics and long-term dependencies, for creating informative summaries. The authors also express an interest in developing more linguistically aware encoders and decoders that incorporate co-reference and entity linking. This research is funded by the European Research Council, the European Union under the Horizon 2020 SUMMA project, and Huawei Technologies. [4] Author: - Shashi Narayan and colleagues focus on extractive summarization, where the goal is to select a fragment of phrases from a document that effectively captures its essential information. Their study proposes a novel approach that utilizes reinforcement learning to train a neural model for ranking sentences, aiming to generate high-quality summaries. The proposed algorithm optimizes the ROUGE evaluation metric through machine learning-based reinforcement learning, globally optimizing the ranking of sentences. This means that sentences are prioritized highly if they appear in summaries with high scores, resulting in the generation of more informative and coherent summaries. The research utilizes various datasets to train the neural summarization model and demonstrates its superiority over state-of-the-art approaches. Results indicate that the proposed approach outperforms related extractive systems across different datasets and metrics, highlighting its effectiveness in producing superior extractive summaries. [5] Author: - josh paul and colleagues introduce a new deep reinforced model for abstractive summarization, aiming to generate natural language summaries while retaining the essential content of input documents. The model incorporates a unique intra-attention mechanism that attends separately to both the input and the continually generated output. Additionally, it utilizes a novel training approach that combines traditional supervised word prediction with machine learning reinforcement. The performance of the model is evaluated using the CNN/Daily Mail and New York Times datasets, achieving a notable ROUGE-1 score of 41.16 on the CNN/Daily Mail dataset, surpassing prior state-of-the-art models. Human evaluation further confirms the model's ability to produce higher-quality summaries. The paper also explores the distinction between extractive and abstractive summarization and provides an insightful summary of related research in the field. [6] Author: - Mathue riya, Abstractive Text Summarization (ATS), two main types are identified: extractive and abstractive. Extractive ATS involves selecting the most crucial sentences from the source document without modification, essentially acting as a binary classification task. Conversely, abstractive ATS generates entirely new sentences to convey the original document's ideas, posing challenges such as preserving essential content and handling long-term dependencies and out-of-vocabulary words. Deep Learning (DL) approaches have gained prominence in abstractive ATS, particularly through deep neural sequence-to-sequence models employing the encoder-decoder architecture. However, this framework may produce trivial summaries, necessitating advancements to ensure the generation of high-quality summaries that maintain the original content's meaning. In addressing these challenges, research has explored Reinforcement Learning (RL) approaches, which have shown promise in various domains, including natural language processing (NLP). RL involves an agent interacting with an environment, learning through trial and error to determine the optimal policy for sequential decision-making. In the context of abstractive ATS, rewards can be developer-defined metrics based on the task, aiming to maximize the summary's coherence and fidelity to the source document's meaning. [7] Author: - Jing and McKeown observed that professional abstractors often reuse text from the original document, adjusting extracted sentences to form the summary. Inspired by this practice, rather than generating new sentences from scratch based on keywords, we adopt a strategy of identifying significant phrases within original sentences (referred to as basic phrases containing keywords) to serve as essential components for an abstractive summary. This approach allows us to minimize the inclusion of ungrammatical phrases and produce sentences closely aligned with the original meaning. To generate a new sentence from the important phrases within the original sentence, we first extract the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

fragment spanning from the first to the last important phrase. This fragment is considered crucial within the original sentence. Subsequently, other words from the original sentence are added at the beginning and end of this fragment to create a syntactically correct sentence. [8] I. F. Moawad and M. Aref [10] introduced a semantic graph reduction approach for abstractive text summarization. Their method involves summarizing a single input document by creating a Rich Semantic Graph (RSG). The approach consists of three main phases: Rich Semantic Graph Creation Phase: This phase begins with a deep syntactic analysis of the input text, followed by the generation of typed dependency relations and syntactic and morphological tags for each word. For each sentence, the model accesses a domain ontology to instantiate, interconnect, and validate the sentence concepts. [11] RSGs represent various semantic representations of the entire document, with the highest ranked RSG considered. Preprocessing Module: This module includes four main processes: named entity recognition, morphological and syntactic analysis, cross-reference resolution, and pronominal resolution. Rich Semantic Sub-Graphs Generation Module: Each pre-processed sentence is represented as a sequence of words (Si = [Wi1, Wi2, … Win]), where each word is represented as a triple sequence (Win = [St, T, D]), with St representing the word stem, T representing the set of tags, and D representing the set of typed dependency relations. Rich Semantic Graph Generation Module: This module generates the final rich semantic graphs from the highest-ranked rich semantic sub-graphs of the document sentences, merging the semantic sub-graphs to form the final rich semantic graph. Rich Semantic Graph Reduction Phase: In this phase, a set of heuristic rules is applied to the generated rich semantic graph to reduce it by merging, deleting, or consolidating graph nodes. Each heuristic rule for a sentence consists of three nodes: Subject Noun (SN) node, Main Verb (MV) node, and Object Noun (ON) node. [9] Author: - Mack henry Prabhudoss Jan Janam and collaborators offer a comprehensive overview of text summarization in their study, with a focus on extracting essential information from large datasets. They trace the evolution of summarization techniques, from linguistic approaches to advanced machine learning methods, covering both single and multi-document summarization strategies. Their study delves into various aspects of summarization, including feature representation, sentence selection, and summary generation, utilizing machine learning, graph-based, and evolutionary approaches. They highlight challenges in automatic natural language processing and discuss recent research articles across different domains performing text summarization. The authors emphasize the effectiveness of dense vector representation and word embedding in addressing issues associated with sparse vectors. [12] They also explore graph-based summarization methods, semantic approaches, and optimization-based techniques, demonstrating their efficacy in producing summaries based on ROUGE scores across various datasets. Furthermore, the study conducts a comparative examination of various summarization models and their applications, offering insights into the evolving landscape of text summarization techniques. It underscores the importance of semantic analysis in summarization and advocates for optimization-based approaches in multi-document summarization. [13] The document presents a range of models and algorithms related to automatic summarization, highlighting the domain independence of summarization tasks and the necessity for improved semantic approaches. Serving as a core study in the field, it provides a comprehensive understanding of summarization techniques and challenges, offering valuable insights for researchers and practitioners in natural language processing. It discusses different learning methods employed in abstractive summarization, showcasing the diversity of approaches and methodologies utilized by researchers in this domain. [14]

3. Motivation

Abstract text summarization is integral to modern information retrieval and natural language processing systems, driven by several compelling motivations. In an age of information overload, where vast data volumes are generated daily, effective summarization techniques are indispensable for distilling crucial information from large text volumes. This allows users to grasp document key points swiftly, saving

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

time and enhancing efficiency. Moreover, abstract text summarization facilitates information dissemination and accessibility, especially in scenarios where time or resources are limited. Concise summaries improve information access for diverse audiences, including those with disabilities or limited literacy, promoting inclusivity and knowledge dissemination. Furthermore, abstract text summarization enhances various natural language processing tasks' performance, like document clustering and categorization. Summarization algorithms aid in identifying core document themes, enhancing related task accuracy and effectiveness. Additionally, abstract text summarization advances intelligent systems and artificial intelligence by enabling machines to comprehend and generate human-like summaries. This technology promotes natural language understanding, fostering more sophisticated AI systems capable of interpreting human language effectively. In conclusion, abstract text summarization revolutionizes textual data interaction and value extraction. By providing concise, accurate, and informative summaries, it empowers decision-making, improves information accessibility, and fosters innovation across industries.

4. Problem Statement

The abstract text summarization techniques have evolved significantly, driven by the need to efficiently handle vast volumes of word-based data. Utilizing pre-trained encoders has emerged as a pivotal method in this domain, enabling the extraction of essential information while catering to user-specific needs. Current literature emphasizes advancements in abstractive summarization, particularly focusing on neural network-based approaches. A comprehensive review of neural network models reveals crucial elements in their design, including encoder-decoder construction, apparatuses, drill strategies, optimization algorithms, dataset selection, and appraisal metrics. This study provides an extensive understanding of recent advancements in neural network-based abstractive summarization models, offering insights into the evolving landscape and highlighting associated challenges. Notably, Bert-based encoder-decoder models have emerged as innovative solutions, representing the forefront of the field. Drawing from this survey, the study advocates for integrating pre-trained verbal mock-ups with neural network techniques for optimal performance in abstractive summarization tasks. Additionally, the paper delves into the historical development of text summarization techniques, from linguistic to advanced machine learning approaches, covering both single and multi-document summarization. It discusses feature representation, sentence selection, and summary generation with machine learning, graph-based, and evolutionary methods, highlighting challenges in automatic natural language processing and recent articles performing text summarization in different domains. Moreover, the study emphasizes the importance of dense vector representation and word embedding to address issues associated with sparse vectors. It discusses graph-based and optimization-based summarization techniques, showcasing the effectiveness of different models based on ROUGE scores for various datasets. The paper serves as a core study of concepts, techniques, and algorithms associated with automatic summarizing, providing a comprehensive understanding of the techniques and challenges in text summarization. It offers valuable insights for researchers and practitioners in natural language processing, highlighting advancements and potential areas for improvement in text summarization techniques. Figure 1 Explains text Pre Processing. [15-18]

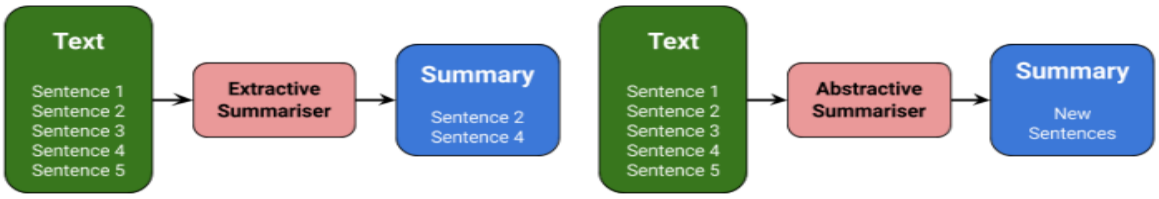


Figure 1 Text Pre-Processing

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

5. Existing Statement

5.1. Introduction

Text summarization, the process of condensing lengthy text into a concise summary focusing on essential information, is gaining traction in various fields. It addresses the time constraints individuals face when consuming articles, research papers, or narratives, allowing for efficient comprehension of main points. Automated text summarization, leveraging Natural Language Processing (NLP) and Artificial Intelligence (AI), offers a cost-effective alternative to manual summarization, albeit facing challenges like sentence ordering and verbosity. However, recent advancements strive for accurate, coherent summaries that cover major points without redundancy. There are two main approaches to text summarization: extractive-based and abstraction-based. Extractive summarization selects important key phrases from the source text to form a summary, while abstraction-based summarization paraphrases the text to create a more semantic representation. Abstractive summarization, although more advanced, is still in early stages of research, while extractive methods remain efficient. Summarization can also be categorized as inductive or informative, with inductive providing the main idea in a brief summary, typically 5% of the source text, and informative offering more detail, around 20%. Various methods, including statistical, machine learning (ML), coherent-based, graph-based, and algebraic-based approaches, have emerged for automated text summarization. ML techniques such as Naïve Bayes, Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are popular, trained to identify sentences likely to form a summary. Pre-processing techniques like tokenization, normalization, and noise removal enhance ML model training. Evaluation metrics like Rouge-N and Rouge-L assess summary correctness. Despite the availability of many ML models for text summarization, research on comparing their performance is limited. Choosing the right ML model is crucial for optimizing training time, reducing machine burden, and improving accuracy. However, extensive text data and research are needed to identify the most effective ML model for text summarization.

5.2. Methodology

The methodology employed in text summarization research involves several key steps aimed at developing and evaluating automated summarization techniques. Here is an overview of the typical methodology: Problem Definition: Define the objectives and scope of the research, including the type of text to be summarized, the target audience, and the desired output format. [19]

- Data Collection: Gather a diverse dataset of text documents representing the type of content the summarization system will handle. This dataset should cover various topics and writing styles to ensure the robustness of the model.
- Pre-processing: Clean and preprocess the text data by removing noise, such as HTML tags or special characters, and tokenizing the text into smaller units, such as words or sentences. Feature Extraction: Extract relevant features from the pre-processed text data, such as word frequency, sentence length, or semantic similarity, to represent the content in a format suitable for machine learning algorithms.
- Model Selection: Choose an appropriate machine learning model for text summarization, considering factors such as the size of the dataset, the complexity of the task, and the computational resources available. Common models include Naïve Bayes, Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF).
- Training: Train the selected model using the pre-processed text data and the extracted features. This involves splitting the dataset into training and validation sets, feeding the data into the model, and optimizing the model parameters to minimize loss and improve performance.
- Evaluation: Evaluate the trained model's performance using standard metrics such as Rouge-N and Rouge-L, which measure the overlap between the generated summaries and reference summaries provided by human annotators.
- Fine-tuning: Fine-tune the model based on the evaluation results to address any deficiencies or

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

limitations identified during the evaluation process. This may involve adjusting hyper parameters, retraining the model with additional data, or incorporating domain-specific knowledge.

- Testing: Test the final model on unseen data to assess its generalization performance and robustness. This step helps ensure that the model can effectively summarize new text inputs outside the training dataset. Comparison: Compare the performance of the developed summarization model with existing baseline models or state-of-the-art techniques to benchmark its effectiveness and identify areas for improvement.
- Analysis: Analyze the results and insights gained from the experimentation process, including the strengths and weaknesses of the proposed methodology, potential future research directions, and practical implications for real-world applications.
- By following this methodology, researchers can systematically develop, evaluate, and improve automated text summarization techniques to address the growing demand for efficient content summarization in various domains.

5.3. System Architecture

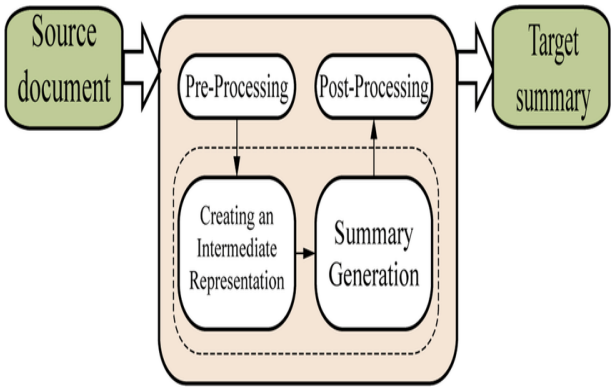


Figure 2 System Architecture

As shown in Figure 2 The system architecture for text summarization involves designing a robust framework that integrates various components to process input text and generate concise summaries. Here's an overview of the typical system architecture:

5.3.1. Input Module:

This module receives the input text data from external sources such as articles, documents, or web pages. It preprocesses the input text to remove noise, including special characters, HTML tags, and formatting inconsistencies. Text Processing Module: The processed text is tokenized into smaller units, such as words or sentences, to facilitate further analysis. This module may include techniques for stemming, lemmatization, and part-of-speech tagging to normalize the text and enhance semantic understanding. Feature Extraction Module: Relevant features are extracted from the tokenized text to represent the content in a format suitable for machine learning algorithms. Feature extraction techniques may include word frequency analysis, TF-IDF (Term Frequency-Inverse Document Frequency), word embeddings, or semantic similarity measures.

5.3.2. Summarization Model:

The summarization model utilizes machine learning algorithms to generate concise summaries from the extracted features. Common models include extractive summarization algorithms, which select important sentences or phrases from the input text, and abstractive summarization algorithms, which generate new phrases to capture the main ideas. The model may be trained using supervised learning with labelled data or unsupervised learning techniques for automatic summarization without human-generated summaries.

5.3.3. Output Generation Module:

The generated summary is formatted and presented to the user in a readable format, such as plain text, HTML, or a graphical user interface (GUI). This module may also include post-processing steps to enhance the coherence and readability of the summary, such as sentence reordering or grammar correction. Evaluation Module: The quality of the generated summaries is evaluated using standard metrics such as Rouge-N (Recall-Oriented Understudy for Gusting Evaluation) and Rouge-L. Human annotators may also provide subjective evaluations to assess the readability, relevance, and overall effectiveness of the summaries. Feedback Mechanism: A feedback loop allows users to provide feedback on the generated summaries, which can be

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

used to improve the performance of the summarization model over time. User feedback may include ratings, comments, or corrections to help refine the summarization algorithm and enhance user satisfaction. [20]

6. Proposed System

6.1. Introduction

The burgeoning availability of online information has necessitated extensive research in automatic text summarization within the Natural Language Processing (NLP) community. Over the past five decades, researchers have tackled this challenge from various angles, spanning different domains and employing diverse paradigms. This survey aims to explore some of the most pertinent approaches in both single-document and multiple-document summarization, with a particular emphasis on empirical methods and extractive techniques. Additionally, the survey investigates promising strategies that target specific aspects of the summarization task. Special attention is devoted to the automatic evaluation of summarization systems, as progress in this area is crucial for future research endeavours. By providing a comprehensive overview of recent advancements in abstractive summarization models, this study sheds light on the evolving landscape and associated challenges. Notably, it highlights the emergence of Bert-based encoder-decoder models as a forefront innovation in the field. Drawing insights from this survey, the study advocates for the integration of pre-trained language models with neural network techniques for optimal performance in abstractive summarization tasks. Keywords: Automatic text summarization, neural network models, abstractive summarization, Bert-based encoder-decoder models, pre-trained language models, empirical methods, extractive techniques, automatic evaluation, summarization systems.

6.2. Methodology

- Literature Search: A comprehensive search was conducted across various academic databases, including but not limited to PubMed, IEEE Xplore, ACM Digital Library, and Google Scholar. Keywords such as "automatic text summarization," "neural network models," "abstractive summarization," and "extractive techniques" were used to identify relevant studies. Figure 3 shows the Flow Cycle.
- Inclusion Criteria: Studies were included if they focused on automatic text summarization techniques, particularly those employing neural network models for abstractive summarization or extractive techniques. Only papers published in peer-reviewed journals or presented at reputable conferences were considered.
- Exclusion Criteria: Studies were excluded if they were not directly related to automatic text summarization or if they lacked sufficient detail on the methodology employed. Non-English articles were also excluded from the review.
- Data Extraction: Relevant data from selected studies, including the author(s), publication year, methodology description, key findings, and limitations, were extracted and organized for analysis.
- Synthesis and Analysis: The extracted data were synthesized and analysed to identify common themes, trends, and gaps in the existing literature. Special attention was paid to the methodology sections of each study to understand the techniques employed and their effectiveness in automatic text summarization.
- Quality Assessment: The quality of included studies was assessed based on established criteria such as study design, methodology clarity, sample size, and rigor of analysis. Studies with methodological limitations were critically evaluated and their impact on the overall findings was considered.
- Results Presentation: The findings of the literature review were synthesized and presented in a coherent manner, highlighting key methodologies, advancements, challenges, and future directions in the field of automatic text summarization.
- By following this methodology, the study aims to provide a comprehensive overview of existing approaches to automatic text summarization, identify areas for further research, and contribute to the advancement of this important area within natural language processing.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

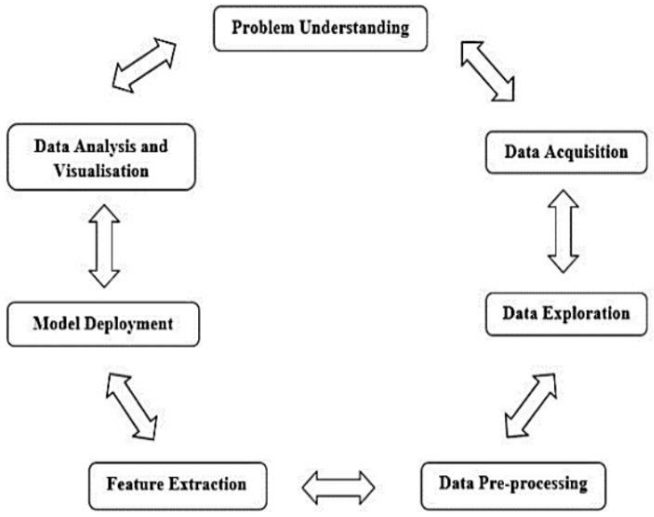


Figure 3 Flow Cycle

Text summarization in machine learning (ML) is often framed as a classification problem, where sentences are classified as either summary sentences or non-summary sentences based on their features. A dataset comprising news articles from CNN and Daily Mail websites was collected and pre-processed. The original text in story format was converted into a tabulated format and stored as a .CSV file. During pre-processing, word tokenization was performed to convert sentences into word-level representations, and redundancies were removed. Following pre-processing, sentences were labelled as positive (summary sentences) or negative (non-summary sentences). Feature extraction from the processed dataset was carried out using word-based TF-IDF (Term Frequency-Inverse Document Frequency). Subsequently, various machine learning algorithms including Support Vector Machines (SVM), Naïve Bayes, Decision Trees, and Random Forest were employed to train the text summarization model. The performance of these models was evaluated using the F1 score metric with validation data. The best-performing model, determined by the F1 score, was selected as the champion model. This champion model was then used to extract summary sentences from the original news stories. The efficacy of the model was further assessed using test data, with the Rouge score serving as a metric to evaluate how well the model performed in generating summaries. This methodology allows for the efficient extraction of summary sentences from news articles, aiding in the summarization process while ensuring that the generated summaries accurately capture the essence of the original content.

6.3. System Architecture

As shown in Figure 4, Some systems employ hybrid approaches that combine both extractive and abstractive summarization techniques with the use of domain-specific knowledge graphs. Extractive methods may utilize graph-based ranking algorithms to select important sentences or phrases, while abstractive methods can generate summaries enriched with domain-specific terminology and context. To further enhance the summarization quality, the system can incorporate a feedback loop mechanism where users provide feedback on the generated summaries. This feedback is used to fine-tune the model parameters, update the knowledge graph, or improve the summarization algorithm iteratively.

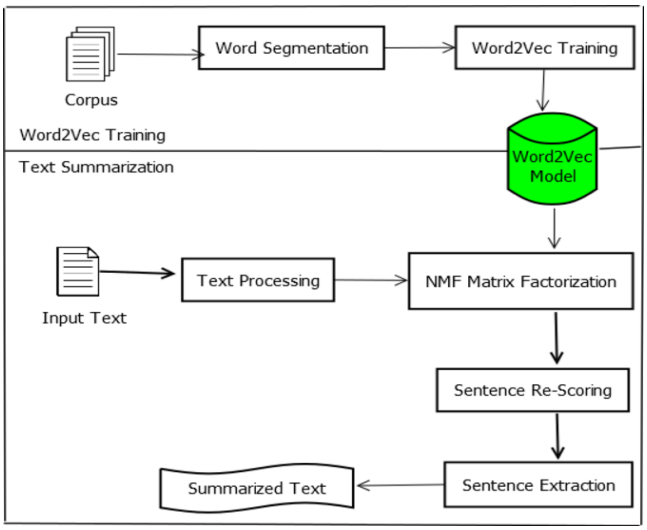


Figure 4 System Architecture

The system architecture described in the provided content involves the utilization of pre-trained encoders for text summarization, particularly focusing on abstractive summarization using neural network-based approaches. The architecture encompasses several key elements: Encoder-Decoder Construction: The architecture employs encoder-

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

decoder models, with emphasis on neural network-based approaches. These models are designed to encode input text into a fixed-length vector representation (encoder) and decode it into a summarized form (decoder). Apparatuses: Various neural network architectures are utilized as part of the encoder-decoder framework. These may include attention mechanisms, recurrent neural networks (RNNs), convolutional neural networks (CNNs), or transformer-based models. Drill Strategies and Optimization Algorithms: The system incorporates drill strategies and optimization algorithms tailored for text summarization tasks. This may involve techniques such as attention mechanisms, sequence-to-sequence learning, reinforcement learning, and specialized optimization algorithms to improve summarization quality. Dataset Selection: The architecture relies on specific datasets for training and evaluation purposes. These datasets may include various corpora, such as news articles, scientific papers, or other text sources, depending on the target domain of summarization. Appraisal Metrics: The system employs evaluation metrics to assess the performance of the summarization models. Common metrics include ROUGE scores, BERT scores, factual scores, and other metrics tailored for summarization tasks. Overall, the system architecture leverages pre-trained encoders, neural network-based models, and specialized optimization techniques to perform abstractive text summarization. It emphasizes the importance of dataset selection, model design, and evaluation metrics in enhancing the summarization process.

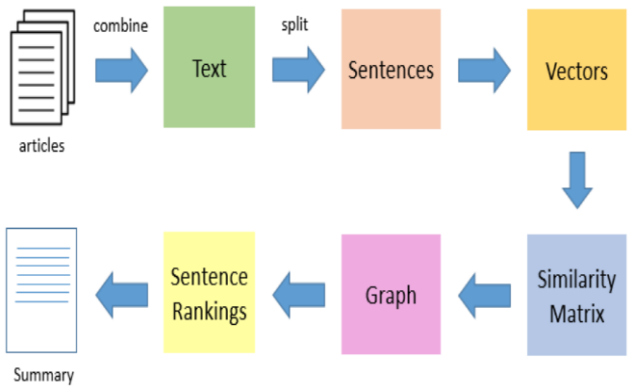


Figure 5 Text Summarization

Text summarization is a critical component of Natural Language Processing (NLP), enabling the condensation of lengthy text data into concise summaries. In essence, it involves distilling voluminous text, sourced from articles, magazines, or social media, into brief, semantic representations. This process is invaluable when time is limited, allowing individuals to obtain the essence of the text without sifting through extensive content. By eliminating extraneous details and retaining only the most pertinent information, text summarization facilitates efficient consumption of information. Furthermore, it aids in comprehension by presenting key insights in a condensed format. Notably, ensuring the originality of the summarized content is paramount to uphold ethical standards and avoid plagiarism concerns.

6.3.1. Extractive Approaches

Using an extractive approach, we summarize our text on the basis of simple and traditional algorithms. For example, when we want to summarize our text on the basis of the frequency method, we store all the important words and frequency of all those words in the dictionary. On the basis of high frequency words, we store the sentences containing that word in our final summary. This means the words which are in our summary confirm that they are part of the given text as shown in Figure 5.

6.3.2. Abstractive Approaches

An abstractive approach is more advanced. On the basis of time requirements, we exchange some sentences for smaller sentences with the same semantic approaches of our text data. The project scope for abstract text summarization includes developing a system that can automatically generate concise and coherent summaries of text documents. This system will utilize algorithms for extractive or abstractive summarization, considering factors such as key phrase extraction, sentence importance, and coherence. The algorithms will be designed to be adaptable to different languages and domains, accommodating variations in writing styles, vocabularies, and document structures. The project will involve defining metrics and benchmarks for evaluating the performance of the summarization algorithms, including metrics for coherence,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

informativeness, and fluency. A user-friendly interface will be designed to allow users to input text documents and view the generated summaries. The summarization algorithms will be integrated into existing systems or platforms for broader application and usability. Thorough testing and validation will be conducted to ensure that the summarization algorithms meet the defined metrics and benchmarks. Documentation will be provided for the development process, algorithms, and results, with reports or publications prepared to disseminate findings. Finally, potential areas for future research and development will be identified to enhance the capabilities and performance of abstract text summarization systems.

7. Comparison Information

Evaluating text summarization poses challenges, particularly in discerning key phrases or content that adds value to the summary. The significance of key phrases can vary depending on the context, making it difficult for machines to identify and locate relevant information accurately. Consequently, automatic evaluation measures are essential for ensuring reliable and effective assessment. Upon reviewing previous research papers on text summarization, various methods for measuring summarization have emerged. These evaluation metrics fall into two main categories:

7.1. Extrinsic Evaluation

Extrinsic evaluation assesses the quality of an automatic text summarization (ATS) system by examining its impact on other tasks such as text categorization, information retrieval, and question answering. A good summary is one that enhances these activities. Extrinsic evaluation methods include relevance assessment, which determines if the summary is pertinent to the topic, and reading comprehension, which evaluates if it can answer multiple-choice assessments.

7.2. Intrinsic Evaluation

Intrinsic evaluation gauges the quality of a summary by comparing machine-generated summaries with human-generated ones as shown in Figure 6. Quality and information are two critical factors in judging a summary's effectiveness. Human experts may employ various quality measures, including readability, non-redundancy, structure, coherence, and other metrics like referential clarity, conciseness, focus, and content coverage. Precision, recall, and F-measure are valuable measures for intrinsically evaluating summaries. Researchers must ensure comparability between human-generated and automatically generated summaries. Despite utilizing these evaluation metrics, it's possible for two summaries to produce different evaluation outcomes, even if they are of equal quality. By employing a combination of extrinsic and intrinsic evaluation methods, researchers can comprehensively assess the effectiveness and quality of automatic text summarization systems, ensuring their utility across various applications and contexts.

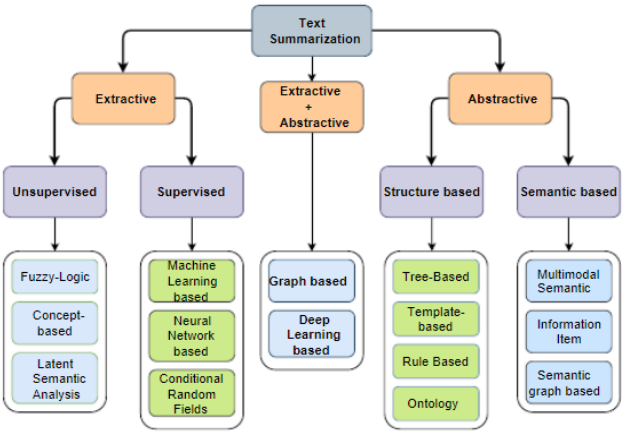


Figure 6 Comparison Information

8. Result and Discussion

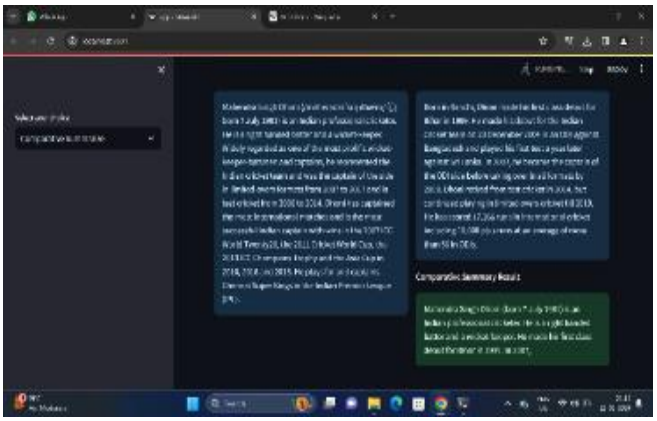
The purpose of the project on abstract text summarization is to develop an automated system that can generate concise and coherent summaries of text documents. This system aims to address the challenge of information overload by condensing 1large volumes of text into shorter summaries that retain the key information and meaning of the original text. The project seeks to improve information retrieval and processing by providing users with quick access to important information without the need to read through entire documents. This is particularly valuable in scenarios where individuals have limited time or resources, as well as in applications where summarization can enhance

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

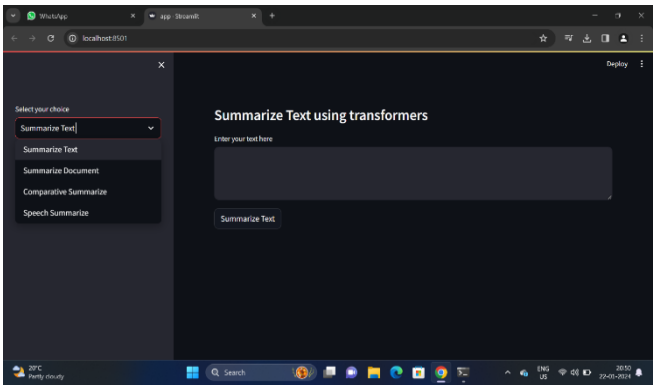
accessibility for individuals with disabilities or limited literacy levels. Furthermore, the project aims to advance the field of natural language processing by developing algorithms and techniques that can accurately and effectively summarize text across different languages and domains. This includes addressing challenges such as variations in writing styles, vocabularies, and document structures. Overall, the project's purpose is to develop a robust and scalable system for abstract text summarization that can be applied in various applications and domains, ultimately enhancing information retrieval, comprehension, and decision-making processes.

8.1. Result

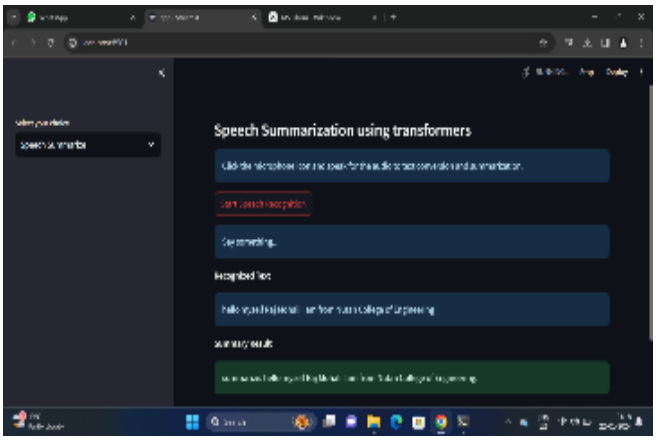
The outcome for this result is vIn the process of creating the summary, several steps are involved. Firstly, raw data input is gathered, followed by text preprocessing to generate organized and cleaned words. Finally, the cleaned text undergoes summarization using TFRSP and a machine learning algorithm. A. Dataset Collection: There are two main methods for acquiring data sources. One approach involves obtaining data in a standard format, such as a .csv file from platforms like Kaggle or other data gathering websites. Another method involves content scraping, where raw data is extracted directly from websites. In this instance, Kaggle is utilized to acquire the Amazon product review dataset, which is then compared with a dataset containing shoe reviews. Following dataset acquisition, preprocessing is conducted, and the summary is generated through a combination of unsupervised and supervised techniques. B. Pre-processing steps: The raw input dataset undergoes several preprocessing stages. Initially, noisy input is examined for duplicates and null values, which are then removed. Subsequently, the records are tokenized, breaking down the text document into sentences, and further into words. Stop words, numbers, punctuation, and special symbols are eliminated from the word collection. The extracted keywords are then lemmatized to determine their root words. Additionally, to ensure uniformity, all uppercase letters are converted to lowercase. Finally, the text is cleaned up, and the words are converted to lowercase for consistency.



7(a)



7(b)



7(c)

Figure 7 (a, b, c) Results

In the process of abstractive summarization using the LSTM approach, two key methods are employed: the encoder and the decoder. The encoder is responsible for encoding the input data and maintaining it in a hidden state, while the decoder decodes this encoded information to generate the precise summary. To

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

facilitate this process, the necessary packages and libraries for abstractive summarization are imported. These packages are crucial for implementing the LSTM approach effectively. Additionally, any required packages are downloaded as per the requirements. Once the necessary setup is completed, the input text to be summarized is provided. This input text undergoes preprocessing steps, including the removal of stop words. Stop words are commonly occurring words in a language (e.g., "the," "is," "and") that often do not carry significant meaning in the context of summarization. Subsequently, a frequency table is created. This table counts the occurrences of each word in the input text. By analyzing this frequency table. This information is valuable for determining the relevance of words in the summarization process. The Results of the system is shown in Figure 7 (a,b,c).

Conclusion

In conclusion, the survey on Abstract Text Paragraph Summarization sheds light on the pivotal role of summarization techniques in handling vast volumes of textual data efficiently. The exploration primarily focuses on recent advancements in abstractive summarization, with a particular emphasis on neural network-based approaches. Key elements such as encoder-decoder architecture, training strategies, dataset selection, and evaluation metrics are thoroughly examined to provide a comprehensive understanding of the landscape. The study highlights the growing importance of pre-trained language models, such as BERT, in enhancing the capability of generating summaries from text. It underscores the distinction between extractive and abstractive summarization paradigms, emphasizing the need for deeper language comprehension in abstractive approaches. Throughout the survey, various methodologies and models proposed by researchers are discussed, along with their contributions to the field. From attentional encoder-decoder RNNs to novel neural architectures, each model aims to address specific challenges in summarization, such as factual correctness, content relevance, and handling out-of-vocabulary terms. Moreover, the survey provides insights into the evaluation metrics used to assess the quality of summaries, encompassing both extrinsic and intrinsic evaluation approaches. It underscores the importance of human evaluation in validating machine-generated summaries and ensuring their coherence, readability, and factual accuracy. Overall, the survey advocates for the integration of pre-trained language models and neural network techniques to achieve optimal performance in abstractive summarization tasks. By fostering a deeper understanding of the evolving landscape and associated challenges, the survey paves the way for future research endeavors aimed at advancing the state-of-the-art in text summarization techniques.

References

[1]. Tomas Mikolov, Ilya Subsieve, Kai Chen, Greg Corrado, Jeffrey Dean," Distributed Representations of Words and Phrases and their Compositionality," arXiv:1310.4546v1 [cs.CL], 2013.
[2]. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space,"arXiv:1301.3781v3 [cs.CL], 2013.
[3]. Kyunghyang Cho, Bart van Marienburg, Dzmitry Bandana, Yoshua Bengio, "On the Properties of Neural Machine Translation: Encoder Decoder Approaches," arXiv:1409.1259v2 [cs.CL], 2014.
[4]. S. Hochreiter and J. Urgent Schmid Huber, "Long Short-Term Memory," Neural Compute., vol. 9, no. 8, pp. 1735–1780, 1997.
[5]. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv:1412.3555v1 [cs.NE], 2014.
[6]. R. Shu,"Residual Stacking of RNNs for Neural Machine Translation", 3rd Workshop on Asian Translation, japan pp. 223–229, 2016.
[7]. D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv:1409.0473v7 [cs.CL] 2016.
[8]. R. Paulus, C. Xiong, and R. Socher, "A Deep Reinforced Model for Abstractive

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0323
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2217-2231

Summarization," arXiv:1705.04304v3 [cs.CL], 2017.
[9]. W. Zeng, W. Luo, S. Fidler, and R. Urtasun, "Summarization with read - again and copy mechanism," pp. 1–11, 2017.
[10]. D. Britz, A. Goldie, M. Luong, and Q. Le, "Massive Exploration of Neural Machine Translation Architectures," arXiv:1703.03906v2[cs.CL], 2017.
[11]. Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, Noam Shazeer, "Generating Wikipedia by Summarizing Long Sequences," arXiv:1801.10198v1 [cs.CL], 2018
[12]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, LlionJones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," arXiv:1706.03762v5 [cs.CL], 2017.
[13]. Kukkar, Ashima, Rajni Mohana, and Yugal Kumar. "Does bug report summarization help in enhancing the accuracy of bug severity classification?" Procedia Computer Science 167 (2020): 1345-1353.
[14]. Yang, Cheng-Zen, Cheng-Min Ao, and Yu-Han Chung. "Towards an Improvement of Bug Report Summarization Using Two-Layer Semantic Information." IEICE TRANSACTIONS on Information and Systems 101.7 (2018): 1743-1750.
[15]. H. Jiang, X. Li, Z. Ren, J. Xuan and Z. Jin, "Toward Better Summarizing Bug Reports with Crowdsourcing Elicited Attributes," in IEEE Transactions on Reliability, vol. 68, no. 1, pp. 2-22, March 2019, doi: 10.1109/TR.2018.2873427.
[16]. Galappaththi, Akalanka. Automatic sentence annotation for more useful bug report summarization. Diss. Lethbridge, Alta.: University of Lethbridge, Department of Mathematics and Computer Science, 2020.
[17]. Kim, Beomjun, Sungwon Kang, and Seonah Lee. "A Weighted PageRank-Based Bug Report Summarization Method Using Bug Report Relationships." Applied Sciences 9.24 (2019): 5427.
[18]. Jiang, He, et al. "Toward better summarizing bug reports with crowdsourcing elicited attributes." IEEE Transactions on Reliability 68.1 (2018): 2-22.
[19]. Bhatia, Surbhi. "A Comparative Study of Opinion Summarization Techniques." IEEE Transactions on Computational Social Systems (2020).
[20]. Ding, Jianli, et al. "Generative Text Summary Based on Enhanced Semantic Attention and Gain-Benefit Gate." IEEE Access 8 (2020): 92659-92668.