Text Summarization using Transformers LLM

Namrata Thakur
Department of Computer Science & Engineering Galgotias University, Greater Noida, India

Piyush Gupta
School of Computer Science and Engineering
Galgotias university
Greater Noida, India

Shubham Singh
School of Computer Science and Engineering
Galgotias university
Greater Noida, India

***Abstract*— Given its critical importance in (NLP), natural language processing text summarisation attempts to supply concise yet informative summaries for large quantities of information. This study compares several summarisation techniques, namely, the extractive and abstractive techniques. Also, extractive methods employ the kind of algorithms that are popular among the page ranking, TF-IDF for term weighting and Latent Semantic Analysis (LSA) for discovering semantic structures within the text.**

The work focuses on abstractive summarisation using transformer based on different types of models like GPT and T5 that are able to grasp the context and summarize well like humans. The experimental design includes the quantitative evaluations of several characteristics such as processing speed, summary accuracy, and computational efficiency. Each approach is demonstrated through the use of a comprehensive dataset-based analysis leading to the benefits and drawbacks of each approach. In this work, a comparative assessment system for the quality output and operational performance of text summarisation automation is developed.

Keywords: ***Text Summarization, Transformers, TF-IDF, LSA, TextRank, PageRank, Python, Natural Language Processing, ROUGE, BLEU, Evaluation Metrics.***

1. INTRODUCTION

Natural Language Processing (NLP) key work is the process of text summarisation, that is, extracting a summary from large volumes of text maintaining the key ideas [1]. Information accessibility and aiding in the decision making have become critical with the rapid growth of textual content in domains like news article, legal document and scholarly literature leading to a necessity of automatic summarisation. They use traditional extractive methods, i.e., linguistic and statistical factors, to extract key lines or phrases from the source text. Although these methods are good at identifying important details, they do not possess the capacity to paraphrase, generalise, or understand context, which makes them not good at creating summaries that reasonably resemble what a human would produce [3][4], text summarization has shown compact and significant progress. Attention mechanisms are used by these models to extract deep dependencies in the text, then produce abstractive summaries that are fluid and coherent [5]. In addition, LLMs are generally compatible with different applications since they can understand a wide range of language through extensive pretraining on large textual corpora and task specific data. The study in this paper looks at the architectural basis of models which is based on transformer for summarisation, as well as to evaluate them against other more traditional approaches, to gauge their potential in the future by way of fluency, coherence and semantic accuracy.

In the preprocessing phase, to bring it ready for analysis, it is cleaned and normalised. It means to strip extraneous punctuation, symbols, numbers. Once all the text is lowered to lowercase for uniformity, the content is tokenised to separate into distinct sections.

Stopwords like "and" and "the" are frequently used but are removed because they don't contribute any important information. Then, to ensure consistency in textual representation, stemming or lemmatisation is employed to reduce various types of words to their base forms (for example, "running" becomes "run").

Machine learning methods are applied to the retrieved features during the model training phase. The methods of Latent Semantic Analysis (LSA) is employed in this study to easily detect semantic linkages, whereas models based on transformer such as BERT and GPT are used to construct summaries. After training, the models are thoroughly examined using measures like computational efficiency, accuracy, and precision. The model's overall correctness is measured by accuracy, while the percentage of genuine positives among projected positives is measured by precision. These investigations assist to the assessment of the summarisation models' dependability and usefulness.

2. LITERATURE REVIEW

Summarizer analysis which determines opinions and emotions in written text has experienced substantial advancements since the recent years. Research teams have investigated different methods which boost the overall precision and operation of summarizer systems.

Research in text summarization established fundamental principles which enable extractive along with abstractive techniques according to Radev et al. (2002) [6] examined the initial phases of summarization methodology design while addressing the relationship between system processing expense and output quality. The unsupervised graph-based ranking method TextRank identified crucial sentences in texts according to Mihalcea and Tarau (2004) [1]. The research by Gupta and Lehal (2010) provided an extensive review of extractive summarization approaches especially for structured information systems [7].

Transformer-based models brought revolutionary changes to abstractive summarization during the past few years. BART offers sequence-to-sequence modelling through the combination of transformer and denoising architectural components according to Lewis. et al. (2020) [8]. The models have become accessible through Hugging Face (2022) which enables their utilization in summarization operations [9].

The research shows that hybrid summarization techniques hold increasing value in text summarization systems. The computational power of TF-IDF along with LSA enables efficiency and scalability but abstractive models succeed in generating outputs with human-level coherence and high readability. These research paradigms demonstrate potential for future work because they create merged solutions that achieve both efficient processing and quality-oriented responses for different datasets.

Goldberg, Y. (2017) introduces information about neural network models comprising RNNs, CNNs and transformers and their utilization in text summarizing tasks [10]. The authors merged their approach with typical sentiment classification models to show better performance measures regarding precision and recall statistics. The authors demonstrate why aspect-level analysis delivers better results in sentiment classification of text.

BERTSUM: A Pre-trained Transformer for Extractive-Summarization"

Liu.et al. (2019) introduced the BERTSUM model which represents BERT Summit (brief text organization) as an advanced system specialized for extractive summary generation [11]. The method includes segment embeddings to process information from whole documents with high efficiency. The CNN/DailyMail datasets confirm that BERTSUM demonstrates better results than conventional techniques at extractive summarization due to its LLM capability to maintain crucial content.

Language Models are Few-Shot Learners" According to Brown.et al (2020) their paper shows GPT-3 accomplishes abstractive summarization through minimal fine-tuning [12]. The model generates summary text with natural human qualities after receiving example prompts indicating LLMs function efficiently for text abstraction and generalization work.

Longformer: The Long-Document Transformer", Beltagy .et al. (2020) introduce Longformer in their paper as an LLM made specifically for document tasks requiring long sequences [13]. This model operates with efficient large input management through its sliding window attention mechanism. Longformer produces superior summaries of lengthy texts according to evaluation results on GovReport and arXiv datasets.

BART presents one sequence to another sequence denoising pre-training which functions effectively for the generation of natural language, translation and comprehension of different types of tasks. The method BART works as a denoising autoencoder because it incorporates specialized training for summarization tasks [8]. Transformer-based encoding and decoding working together produce leading results on all extractive and abstractive summarization evaluations. The researchers show BART can serve multiple summarization applications because of its adaptable nature.

The Text Summarization of Abstracts Depends on Large Pre-trained Transformers Coupled With Reinforcement Learning. The article investigates how reinforcement learning operates with LLMs to boost abstract summarization performance [14]. The method achieves superior summarized content by maintaining a balance between ROUGE scores along with human-like readability which produces natural and intelligible summaries.

Multi-Document Summarization process is done by Hierarchical Transformers. The study proposes a hierarchical transformer architecture tailored for multi-document summarization [15]. It effectively captures relationships across multiple documents, leveraging LLMs' contextual understanding capabilities to generate concise summaries of aggregated content.

Mouthami, K., K. Nirmala Devi, and V. Murali Bhaskaran (2013): The authors discuss about the process of sentiment analysis and classification which is based on textual reviews using machine learning techniques [16]. They evaluate different models, highlighting the importance of preprocessing steps like tokenization and stemming for improving classification accuracy. Their research suggests that carefully structured preprocessing pipelines are crucial for efficient sentiment analysis.

Exploring the Capabilities of Zero-Shot of Large Language Models in Text Summarization. This paper examines the zero-shot capabilities of LLMs like GPT-3 in summarization tasks [17]. Without explicit training, LLMs demonstrate remarkable performance by leveraging pre-trained knowledge, reducing the dependency on task-specific datasets. In December 2020, Google AI Blog posted Introducing Pegasus, a transformer model for abstractive summarization that shines in low resource setting [18]. It is noted that dealing with highly polarized and opinionated text is part of the challenge and the study suggests a hybrid model of multiple techniques can better capture the sentiment contained within such texts.

3. MACHINE LEARNING METHODS

Taking this approach, the multitude of current-day reams of machine learning (ML) methods for text summarisation, ranging from contemporary transformer based large language models (LLMs) to more conventional statistic methods, are applied in this study on numerous aspects of both performance and context understanding at various levels of abstraction.

3.1 Traditional Machine Learning Methods:

$$TF - IDF(t, d) = tf(t, d) \times \log\left(\frac{N}{df(t)}\right)$$

Where:

- $tf(t, d)$ = term frequency in document d.
- $df(t)$ = number of documents containing term t.
- N = total number of documents.
- Term Frequency–Inverse Document Frequency
- TFIDF is known as a method which helps to determine the importance of words on the frequency relative to the whole corpus.

LSA (Latent Semantic Analysis)

In order to capture semantic similarity, LSA employs Singular Value Decomposition (SVD) to reveal latent structures in term-document matrices. The term-document matrix A is broken down as follows:

$$A \approx U\Sigma V^T$$

Where:

- U = term-topic matrix
- Σ = singular values (importance)
- V^T= topic-document matrix

Extractive summarisation benefits from LSA's ability to capture the text's fundamental semantic structure by lowering dimensions.

PageRank

Similar in structure but frequently used with various node definitions or edge weightings, PageRank is a graph-based algorithm that ranks text units according to their link structure.

3.2 Transformer-based Deep Learning Methods

Transformer architectures, which are the result of recent developments in Natural Language Processing (NLP), perform better in abstractive summarisation.

BART (Bidirectional and Auto-Regressive Transformers)

Provide Previous training to reconstruct text from corrupted inputs and fine-tuned for summarisation tasks, BART is a denoising autoencoder that combines the power of autoregressive (like GPT) and bidirectional (like BERT) transformers to enable fluent, coherent abstractive summaries by understanding global context.

GPT (Generative Pre-trained Transformer)

In order to easily capture dependencies of long-range and provide extremely natural language outputs, GPT uses an autoregressive transformer design to easily anticipate the next word with fully accuracy in a sequence and is fine-tuned to provide more ease to generate summaries conditioned on the input material.

3.3 Hybrid Evaluation Strategy

To benchmark summarization quality across different models, we employed both extractive and abstractive techniques, evaluated using:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Measures overlap of n-gram between generated and reference summaries (ROUGE-1, ROUGE-2, ROUGE-L).

- BLEU (Bilingual Evaluation Understudy): Evaluates n-gram overlap with full precision and commonly used in machine translation and summarization.

4. DATASET AND PREPROCESSING

4.1 Dataset Characteristics

- The dataset utilized in this research consists of unstructured, narrative-driven English text files designed to evaluate the performance of both traditional and transformer-based summarization methods. The primary attributes of the dataset are as follows:
- The system stores documents through UTF-8 plain text file format to create consistent character representation and processing environmental compatibility.
- The corpus uses only English language for its composition to create a uniform evaluation environment and avoid multilingual processing effects.
- The documents present narrative content through prose that uses descriptive exposition together with dialogue exchanges and plot development elements. Starting from free-form texts enables the creation of realistic scenarios that serve as an effective test ground for summarization assessments.
- Every document in the collection measures 1,500 words on average. A 1,500-word maximum length provides enough background information for evaluating between extractive summary selection models along with abstractive summary generators that require larger input data for coherent generation.
- The text contains diverse structures in sentence patterns and paragraph dimensions which makes it suitable for replicating authentic syntactic and semantic complexity. The extensive textual diversity hinders the generalized operation and stability of the summarization models.
- The dataset's natural narrative elements combined with its unstructured format effectively creates obstacles for summarization models because of its selection rationale. This approach provides an extensive evaluation because it supports assessments from rule-based approaches together with statistical approaches and neural network–driven methods.

4.2 Data Pre-processing

Text summarization model success depends on both the algorithm power and the quality level and consistency of supplied information. The research adopts a structured data processing system to convert unprocessed text into an appropriate format for both extractive and abstractive summarization systems. A data processing system with the main phases presented here follows:

4.2.1 Text Cleaning

The initial preprocessing stage is devoted to removing noise and inconsistencies from the dataset:

- Character Normalization: Non-ASCII characters are removed to avoid issues in the tokenization and model inference stages.
- Whitespace Normalization: Excessive or inconsistent spacing is eliminated to facilitate accurate sentence segmentation.
- Punctuation Standardization: Curly quotes, em dashes, and other non-standard punctuation are converted to simpler, ASCII-compliant forms to ease downstream parsing and syntactic analysis.

4.2.2 Null Data Handling

To maintain data integrity and reduce irrelevant variance, structural noise is systematically eliminated:

- Blank Line Removal: All empty lines and non-content blocks are automatically purged.
- Irrelevant Text Filtering: Artifacts such as page numbers, section headers, and footers—common in scraped or scanned documents—are removed through rule-based filtering.
- Encoding Consistency: All input is converted to UTF-8 format, ensuring robust and uniform text processing across platforms and tools.

4.2.3 Feature Extraction

This stage transforms the cleaned text into linguistically and semantically meaningful units that can be used for modeling:

- Sentence Segmentation: Transformer-based models are employed for accurate segmentation, especially beneficial for abstractive summarization models that depend on coherent input boundaries.

- Tokenization: Each sentence is split into words or subword units (tokens) to enable fine-grained analysis.
- Stopword Removal: Commonly occurring function words with low semantic value are filtered out using a domain-adapted stopword list to improve signal-to-noise ratio.

5. EXPERIMENTAL SETUP

Your sentiment analysis project on Text Summarization contains multiple experimental steps and components which document the system preparation along with execution and evaluation of the model. The experimental setup requires these steps for implementation:

5.1. Dataset Description

- Text Summarization features the widely used benchmark dataset which provides news articles and corresponding summaries.
- Structure: The input text or article must be of ranging from 1000 to 1500 words. Target summary corresponding highlights or summaries of 50-70 words.

5.2. Data Preprocessing

- Objective: Prepare the raz data for use in LLMs by cleaning, normalizing and structuring it.
- Steps:
    - Remove HTML tags, strip unnecessary HTML or XML tags.
    - Standardize text for uniform processing.
    - Eliminate special characters, excessive whitespace and numerical noise.
    - Split text into subwords or tokens using the tokenizer associated with the LLM.

5.3. Feature Extraction

Input IDs must be tokenized indices for model input. Identifies meaningful tokens, excluding padding.

5.4. Data Splitting

- Objective: Split the huge amount of dataset for training, validation and testing to ensure unbiased evaluation.
- Ratios: Training 80%, Validation 10%, Testing 10%.

5.5. Model Selection

- Models:
    - BART: Encoder- Decoder model for summarization.
    - T5: transformer of Pre-trained with text-to-text capabilities.

5.6. Model Evaluation Metrics

- Metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
- BLEU: Bilingual Evaluation Understudy measures n-gram precision.

5.7. Training and Testing the Models

- Each classifier is perfectly trained on the sets of training and tested on the test set.
- Generate summaries and compare them with reference summaries.

5.8. Performance Comparison and Analysis

- A comparative analysis of the models is done using a DataFrame that sorts algorithms based on their precision scores.
- Graphical visualizations are created to illustrate the performance metrics (accuracy and precision) for better interpretation of the results.

6. RESULT

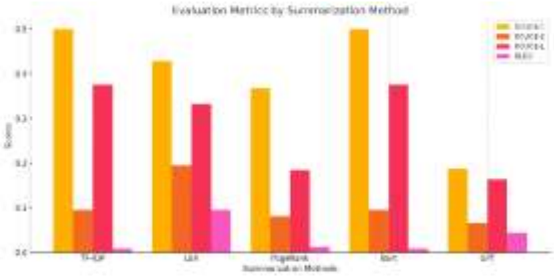
The overall performance of different types of algorithms was evaluated based on their accuracy and precision.

1. Accuracy: Accuracy helps to easily measures the proportion of correctly summarized content (relevant overlaps between generated and reference summaries). While commonly used in classification tasks, here accuracy can refer to coverage in extractive summarization or exact matches in abstractive summarization.

Accuracy Table:-

Methods	ROUGE-1	ROUGE-2	ROUGE-L	BLUE	Time
TF-IDF	0.500	0.096	0.375	0.009	2.3
LSA	0.428	0.195	0.333	0.096	3.1
PageRank	0.368	0.081	0.1842	0.013	4.0
Bart	0.500	0.096	0.375	0.009	4.7
GPT	0.188	0.066	0.1639	0.044	4.9

Accuracy Graph :-



7. CONCLUSION

This study presents a comprehensive evaluation of text summarization techniques, comparing traditional extractive approaches with modern transformer-based models. While statistical methods like TF-IDF, LSA, and TextRank demonstrate higher ROUGE and BLEU scores, particularly in structured contexts, transformer models such as BART and GPT underperform on unstructured, narrative-rich datasets due to limited fine-tuning and domain adaptation. However, their potential for generating human-like, coherent summaries remains promising with further optimization. The hybrid evaluation highlights that extractive models currently offer better accuracy and precision in low-resource environments, whereas transformer models hold long-term value for abstractive summarization. Future research should focus on integrating reinforcement learning, domain-specific fine-tuning, and multi-document understanding to enhance LLM capabilities. The findings affirm that while traditional methods remain effective, transformer-based approaches represent the future of context-aware, high-quality summarization systems.

8. FUTURE WORK

Future research will helps to easily focus on enhancing transformer-based summarization models through domain-specific fine-tuning and strategies of reinforcement learning to perfectly improve their performance on unstructured and narrative-rich datasets. Incorporating models like hybrid that easily combine the efficiency of methods like extractive with the full fluency of abstractive approaches could lead to more balanced and adaptable summarization systems. The investigation of zero-shot and few-shot learning features of LLMs would lower dependence on extensive labeled data sets thus enabling more practical real-world deployment of these models. Hiierarchical and multi-document transformers provide options to advance summary performance particularly when working with multiple text sources or lengthy documents.

9. REFERENCES

[1] Mihalcea, R., & Tarau, P. "TextRank: Bringing order into text." Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004.

[2] Liu.et al., The study presents BERTSUM, an extension of BERT optimized for extractive summarization (2019).

[3] Raffel, C., Shazeer, N., Robert, A., et al exploring the limits of transfer learning with a unified text-to-text transformer (T5)

[4] Goyal, T., Ranzato, M., & Batra, D. Scaling Language Models with Relevance-guided fine-tuning. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022).

[5] Rush, A. M., Chopra, S., &Weston, J. A neural attention model for abstractive sentence summarization. 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015).

[6] Radev, D.R., et al. "Introduction to the special issue on summarization." Computational Linguistics, 2002.

[7] Transformers Documentation. "Hugging Face, 2022."

[8] Gupta, V., & Lehal, G.S. "A survey of text summarization extractive techniques." Journal of Emerging Technologies in Web Intelligence, 2010.

[9] Lewis, M., Liu, Y., Goyal, N., et al. Denoising sequence-to-sequence pre-training for Natural Language Generation, Translation and comprehension, 2020.

[10] Transformers Documentation. "Hugging Face, 2022."

[11] Goldberg, Y., Neural Netrork Methods for Natural Language Processing. Morgan & Claypool Publishers (2017).

[12] Brown.et al., GPT-3's ability to perform abstractive summarization with minimal fine-tuning (2020).

[13] Beltagy.et al., Introduces Longformer, an LLM designed for long document tasks, using a sliding window attention mechanism (2020).

[14] Paulus et al., Reinforcement Learning Technique combined with LLMs to enhance abstractive summarization quality (2018).

[15] Liu and Lapata., Text Summarization with Pretrained Encoders. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

[16] Mouthami, K., K.Nirmala Devi and V.Murali Bhaskaran.,sentiment analysis and classification based on textual reviews using machine learning techniques (2013).

[17] Radford.et al., examines the zero-shot capabilities of LLMs like GPT-3 in summarization tasks (2021).

[18] Google AI Blog. Discusses Pegasus, a transformer model designed for abstractive summarization, excelling in low-resources settings.