









Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction

Deepak Dharrao¹, Manasvi Mishra¹, Aqsa Kazi¹, Madhuri Pangavhane², Priya Pise³,
Anupkumar M. Bongale^{4*}

¹Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune 412115, India

²Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India

³Department of Computer Engineering, Indira College of Engineering and Management, Pune 410506, India

⁴Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune 412115, India

Corresponding Author Email: ambongale@gmail.com

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ria.380311>

ABSTRACT

Received: 25 October 2023
Revised: 27 November 2023
Accepted: 30 December 2023
Available online: 21 June 2024

Keywords:

text summarization, abstractive text summarization, BART, T5, PEGASUS, business news summarization

Nowadays, deep learning models are used to summarize the large volume of text data to understand its intent effectively. Processing huge amounts of data can lead to an Information overload, where the models may generate text summaries that miss out on the important information of actual text content. Such problems in business document summaries can impact progressive business growth. This study employs a dataset comprising business articles sourced from BBC News to conduct an extensive comparative analysis of three prominent text summarization algorithms: Bidirectional and Auto-Regressive Transformers, Text-to-Text Transfer Transformer, and Pre-training with Extracted Gap-sentences for Abstractive Summarization, within the domain of business news summarization. The primary objective is to assess the accuracy of these models in generating concise and coherent summaries, utilizing ROUGE and METEOR scores as the benchmark for evaluation. Each model's proficiency in distilling business narratives while retaining crucial insights is carefully examined. This study analyzes the summaries generated and compares them with the already existing summaries. From the result analysis it observed that BART and PEGASUS show ROUGE-I score of 0.308 and 0.245, and METEOR score 0.28 and 0.25 respectively. The outcomes of this study show that the T5 excelled in the ROUGE-1 and METEOR scores which were 0.354 and 0.35 respectively. Outcomes of this research offer significant implications for both researchers and practitioners, equipping them with advanced summarization techniques for extracting information effectively from business-related content. In an age where information overload is prevalent, the findings from this study can guide the selection and deployment of text summarization models to enhance information extraction processes, ultimately facilitating more efficient decision-making and information dissemination in the business domain.

1. INTRODUCTION

The demand for precise and effective information extraction and condensation techniques has never been greater than it is today, in an era of information overload where massive amounts of text are produced every day throughout the digital landscape. This problem is addressed by text summarizing, a crucial area of natural language processing [1] (NLP), which creates succinct, coherent summaries from extensive texts automatically. In order to facilitate the extraction of important information and the creation of summaries that are human-like, this research project aims to take advantage of the capabilities of cutting-edge pre-trained models, such as BART [2] (Bidirectional and Auto-Regressive Transformers), T5 (Text-to-Text Transfer Transformer) [3], and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive

Summarization) [4].

Information overloading is a prominent and weakly addressed research issue while generating text summaries. Generally, while summarizing the text content, the deep learning models have to refer to enormous corpora of text documents and various text sources. The summarizing model can miss out the important information and may generate unrelated text in the summary. Even the summary could be too short, so essential information in the text summary could be missed. This problem increases when a huge volume of text needs to be processed and summarized. Such issues are critical in business models and can negatively affect the growth of successful business. The described problem is termed as information overloading, and deep learning models are prone to this problem. The need for precise information extraction and condensation techniques has increased to unprecedented

levels in the current era of information overload, where an overwhelming volume of text is generated daily across the digital landscape. This problem is addressed by text summarization, a crucial component of natural language processing (NLP), which automatically generates clear and coherent summaries from lengthy texts [1]. Pre-trained language models are a notable advancement in natural language processing (NLP) that could lead to significant improvements in text summarization automation. With the help of cutting-edge pre-trained models, such as BART (Bidirectional and Auto-Regressive Transformers) [2], T5 (Text-to-Text Transfer Transformer) [3], and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) [4], this research project hopes to produce summaries that are human-like and make it easier to extract important information.

This research investigates the techniques, approaches, and resources used to apply these models for automated text summarization in order to tackle the problem of information overload. Examining the fundamental framework and mathematical foundations of each model provides insight into how they understand and generate textual content. Additionally, the study shows how these models can be used in practice by producing summaries for a range of input articles. Evaluation standards like METEOR [5] and ROUGE [6] are used to rate the quality of the summaries that are produced. The main objective of this research is to automate the summarization process by quickly extracting important insights from the growing body of textual data by utilising the power of pre-trained models.

The use of state-of-the-art pre-trained language models, such as PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization), T5 (Text-to-Text Transfer Transformer), and BART (Bidirectional and Auto-Regressive Transformers), is the main focus of this study. T5 employs a text-to-text framework, BART is proficient in bidirectional and auto-regressive transformations, and PEGASUS creatively employs pre-training with extracted gap-sentences for abstractive summarization.

Each model-the flexible text-to-text framework of T5, the bidirectional approach of BART, and the specific pre-training with extracted gaps of PEGASUS-brings a special strength to the table. Their comparison sheds light on subtle variations in the way these models interpret and handle text, which advances our knowledge of how these models work to automate text summarization.

This paper's later sections go into greater detail about each model, describing how it was used and the results it yielded. This exhaustive investigation establishes the foundation for an in-depth comprehension of how these models work in automating the process of text summarization.

2. RELATED WORK

An essential task in natural language processing known as automated text summarization [7] has attracted a lot of interest lately. There have been several methods investigated, which can be generically divided into extractive and abstractive summarization. While abstractive summarization creates new phrases that capture the essential information, extractive summarization [1, 8] entails choosing and extracting specific sentences from the source text.

The field of text summarization has come a long way from

simple rule-based techniques to the sophisticated state-of-the-art approaches used today. In its early iterations, rule-based summarization extracted important sentences or phrases from a given text by applying predefined grammatical and linguistic rules. By laying down the foundational ideas of information extraction and condensation, this early period prepared the way for later advances.

The development of text summarization techniques reveals an amazing trajectory that reflects the constant quest for better ways to condense information. From the earliest rule-based techniques to the complex, cutting-edge techniques used in modern natural language processing (NLP), this journey can be followed.

Text summarization was first based on crude rule-based techniques, which involved sorting through the text and extracting important sentences or phrases using predetermined grammatical and linguistic rules. By developing the fundamental ideas of information extraction and condensation, this era set the stage for later developments. Rule-based summarization followed preset linguistic structures in an attempt to extract the main ideas of a text.

With the advent of statistical methods, the field experienced a paradigm shift away from strict rule-based approaches. At this stage, algorithms started evaluating a sentence's importance by looking for statistical patterns in the text. During this time, extractive summarization-which involves identifying and presenting existing sentences to compose a concise summary-became increasingly popular. By using statistical analysis to guide sentence selection, this method sought to preserve the most important and instructive passages from the original text.

The introduction of machine learning, and especially the rise of deep learning methods, brought about a paradigm change. The era of abstractive summarization began with this revolution, in which models were able to produce summaries that resembled those of humans by rephrasing and paraphrasing the original text. Abstractive methods, as opposed to extractive summarization, concentrated on comprehending the context and meaning of the content, enabling a more nuanced representation in the summary. This was a big step forward, as models were now able to understand the underlying semantics and produce more concise but still coherent versions of the original content, going beyond simple sentence selection.

The development of text summarization techniques over time can be summarized as a trip from rule-based simplicity to statistical analysis and finally to the revolutionary power of deep learning. The transition from extractive to abstractive summarization is indicative of an ongoing effort to capture the subtle meaning of textual content, which has led to the development of advanced methodologies used in modern NLP. Table 1 shows the summarized literature review of related works.

Abstractive summarization:

Abstractive summarization is the process of generating more human-like, concise summaries using NLP [9]. Text summarization tasks have been very famous due to the fast-moving life in the 21st century. Let's look at the most used approaches:

(1) Transformer-Based Techniques: Transformer-based techniques are the techniques we have focused on in this paper. These techniques have gained popularity because of their ability to produce human-like, accurate, and concise summaries.

(2) Pointer Generator Networks: These networks are known to combine abstractive and extractive techniques to address the difficulties in content selection and fluency. These, however, have a few drawbacks, which include limited semantics understanding cannot handle rare words or new words well.

(3) Graph-Based Approaches: Graph based Neural Networks (GNNs) are used when the document structure and relationships between the sentences are very important.

(4) Reinforcement Learning approaches: These models are re-trained to generate more human-like summaries until we achieve a desired output.

Extractive summarization:

Using key phrases or sentences that are extracted straight from the original content, extractive summarizing techniques create summaries. Below is a quick summary of the extractive summarization techniques that were discussed:

(1) Term Frequency Inverse Document Frequency: The statistical measure known as TF-IDF is used to assess a word's significance in a document in relation to a corpus, or collection of documents. Sentences with higher TF-IDF scores-a measure of their importance in the document-are chosen for the summary in extractive summarization. Easy to understand and effectively computed. It draws attention to terms that are unique to a manuscript. However, it has limited comprehension of semantics; might miss word linkages.

(2) Cluster-Based Method: Sentences with comparable content are grouped into clusters using cluster-based algorithms. Next, sentences that best represent each cluster are chosen to be included in the summary. K-means clustering, and hierarchical clustering are two popular methods. Gather similar sentences to capture thematic consistency. May have trouble with a variety of topics within a document, sensitive to parameter settings and clustering method selection.

(3) Graph Theoretic Approach: Sentences are modelled as nodes in a graph-based approach, while the interactions between sentences are modelled as edges. To find crucial sentences for extraction, centrality metrics like degree centrality and PageRank are employed. Uses graph topology to capture the links and significance of texts. May have trouble with long-range dependencies; sensitive to how relationships are represented.

(4) Latent Semantic Analysis: Singular value decomposition (SDVD) is a technique used in LSA analysis to examine the links between terms and concepts in a document. Sentences that make the most contributions to the underlying latent semantic structure are identified and chosen in extractive summarization. Able to manage polysemy and synonymy; captures semantic links. May have trouble comprehending meanings particular to a certain situation; requires a sizable corpus for training.

Table 1. Displays the literature review of related works

Ref.	Brief	Findings	Inferences
[2]	BART, a denoising autoencoder, is introduced in the article, trained on corrupted text and proficient in reassembling the original content. It extends its capabilities to generalize other pretraining techniques such as BERT and GPT, employing a traditional Transformer-based architecture. Notably, the authors find that the best results come from random phrase shuffling and employing a creative in-filling approach during training.	BART matches RoBERTa in GLUE and SQuAD but shines in text generation, especially abstractive tasks. It outperforms back-translation in machine translation by 1.1 BLEU. Ablation experiments affirm BART's task consistency, highlighting noise reduction improvements and variable control in pretraining.	BART, a denoising autoencoder, uses a Transformer-based architecture to reconstruct text by training with corrupted input. Creative in-filling and shuffled phrase sequences are effective training techniques. BART excels in various NLP tasks, setting state-of-the-art benchmarks in text generation, abstractive discourse, question answering, summarization, and machine translation. Its versatility and performance make it a potent tool in NLP.
[3]	This NLP research introduces a semi-supervised approach to text summarization, inspired by CycleGAN. The model can transfer the style of a document to its summary and vice versa. Applied to Chinese documents using a T5-based model and CSL/LCSTS datasets, it effectively condenses lengthy texts into concise summaries using minimal labeled data, making it practical for real-world applications.	It performs well but lags behind newer fully supervised models. Results support its effectiveness in summarization but note performance gaps and reliance on limited labeled data. Future research should aim to reduce this dependence, improve the architecture, and explore domain adaptability. Datasets, baselines, and algorithms are available for further study.	The article suggests an NLP text summary technique that sees style transfer as a semi-supervised learning job and treats text summarization as such. The technique, which is based on a T5 model, is tested on the CSL and LCSTS datasets in China and shows promise for producing summaries of extensive texts. Although current supervised models still outperform the suggested method, it is more practical and efficient for real-world applications since it makes use of unlabeled data and only needs a small number of labelled samples. Reducing the dependency on labelled data and improving the model architecture can be the main goals of future study.
[4]	The article introduces PEGASUS, a novel abstractive text summarization method using Gap Sentences Generation (GSG) for self-supervised pretraining. GSG involves removing key sentences from input documents and generating summaries from the remaining content. PEGASUS achieved cutting-edge performance across 12	The essay covers PEGASUS design and compares it to BART, T5, and UniLM. Ablation experiments studied pre-training variables and identified the best configurations for the final PEGASUS model. PEGASUS excels in abstractive summarization across disciplines, offering cost-effective summarization potential.	PEGASUS, a novel abstractive text summarization approach, employs the Gap Sentences Generation (GSG) objective for pretraining large Transformer-based models. It achieved cutting-edge performance across all 12 tested summarization tasks, encompassing news, science, stories, instructions, communications, patents, and legislative documents. Remarkably, with just 1000

summarization tasks, including news, science, and legislation. It outperformed prior state-of-the-art results with just 1000 instances in low-resource summarization. Human evaluations confirmed that PEGASUS generated summaries at a human-level quality across various datasets.

The evaluation tool ROUGE (Recall-Oriented Understudy for Gisting Evaluation) presents four distinct methods to automatically assess a summary's quality by contrasting it with other perfect summaries written by humans. These metrics track the number of overlapping units, such as n-grams, word sequences, and word pairs, between the computer-generated summary and the ideal summaries. The metrics have been utilised in extensive summary evaluations and have demonstrated a strong correlation with subjective assessments.

A study compared Bart and T5 models for news article summarization using keywords. Bart slightly outperformed T5 among approximately 1000 articles, especially for mid-sized news pieces. The study also highlighted challenges in news summarization, including manual selection and the presence of ads, which were mitigated using the newspaper3k Python tool. Automated summarization offers significant advantages in delivering article highlights.

Automated evaluation methods like ROUGE, including ROUGE-L, ROUGE-W, and ROUGE-S, are popular due to their efficiency and excel in single document, short, and multi-document summarization across various datasets.

Excluding stopwords often enhances performance in ROUGE, which closely aligns with human judgments, especially in single-document and concise summaries, though multi-document summarization correlations remain a challenge.

Researchers experimented with Bart and T5 models using the BBC News Summary dataset through the Huggingface transformers API.

Bart consistently outperformed T5 in summarization, with higher Rouge scores and an F1 score averaging 33% compared to T5's 26%. Bart's superior recall and precision make it more efficient for medium-sized news article summarization, saving time and resources.

samples, it outperformed prior state-of-the-art results in low-resource summarization. Human evaluations confirmed its summaries matched human-level quality across various datasets. PEGASUS was fine-tuned and compared to pretraining models like BART, T5, and UniLM for optimal performance.

In order to assess a summary's quality by contrasting it with ideal summaries written by humans, the ROUGE assessment package presents four metrics (ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S).

These metrics track the number of overlapping units, such as n-grams, word sequences, and word pairs, between the computer-generated summary and the ideal summaries.

The ROUGE measures have been examined on several datasets, and the results demonstrate good performance in single-document summarization tasks and very short summary tasks, although human judgement correlations in multi-document summarization tasks remain challenging.

When it came to creating summaries of news stories, the Bart model fared better than the T5 model, however the difference was not large.

Challenges in news summary include the need for manual selection of key elements and the existence of irrelevant content and advertisements.

The Bart model was more effective at summarising news because it regularly outperformed the T5 model in terms of summary generation and had higher Rouge scores.

3. PRE-TRAINED LANGUAGE MODELS

This section introduces three top models-BART [2], T5 [3], and PEGASUS [4] in the quickly developing field of natural language processing and automated text summarization [10]. These models are state-of-the-art methods for abstractive summarization, each having an own design and goals. BART excels at producing coherent summaries thanks to its dual nature and encoder-decoder structure. T5's unified text-to-text framework, which offers amazing flexibility, revolutionises NLP [11] tasks. With its unique gap-sentence generating purpose, PEGASUS distinguishes out and excels at producing succinct and insightful summaries. This section offers a succinct summary of these models, emphasising the substantial contributions they have made to automated summarization.

3.1 BART (Bidirectional and Auto-Regressive Transformers)

BART is a transformer-based architecture [2] that has shown success in a variety of natural language processing tasks, including text summarization. It is represented by the model name "facebook/bart-large-cnn." BART demonstrates a dual nature because it incorporates both auto-regressive and bidirectional traits. The approach makes use of an encoder-decoder architecture, which is essential for tasks involving sequences of events, including summarization [12, 13]. The denoising autoencoder objective of BART, which may be mathematically represented as:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N \log p(x_i | y_i; \theta) + \lambda \sum_{i=1}^N \log p(y_i | x_i; \theta) \quad (1)$$

where,

N represents the number of training examples.

x_i is the input text.

y_i is the corresponding target text (summary).

θ denotes the model parameters.

λ is a regularization hyperparameter.

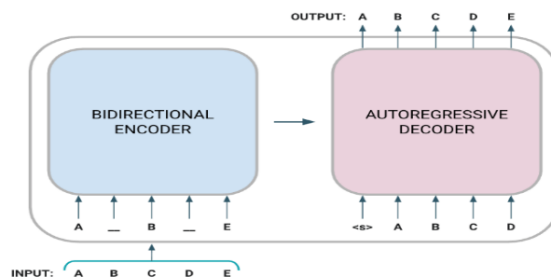


Figure 1. BART model architecture

While the second part of the objective promotes the creation of coherent summaries, the first term pushes the model to reconstruct the input text from corrupted copies. BART is equipped to perform well on tasks requiring abstractive

summarization [14] due to the denoising target and the bidirectional encoding, as shown in Figure 1.

In order to efficiently capture contextual information, BART employs a bidirectional approach to understand and represent input text from both directions. Notably, BART presents an auto-regressive training strategy that enables the creation of coherent and contextually rich abstractive summaries by teaching the model to predict the original sequence from corrupted versions. BART is capable of comprehending and combining textual content for tasks such as text summarization because of its bidirectional architecture and auto-regressive feature.

3.2 Text-to-text transfer transformer (T5)

The T5 [15] framework includes the "t5-small" model, a flexible pre-trained model for comprehending and producing natural language. By using a unified text-to-text framework [15], where both input and output are represented as text, T5 completely transforms how NLP tasks are structured. Specifically using an encoder-decoder configuration for sequence-to-sequence activities, the transformer architecture forms the foundation of T5's fundamental architecture. Despite the fact that T5's architecture is not defined by a particular mathematical equation, the text-to-text strategy guarantees consistency in performing a range of NLP [7] tasks, including summarization [16, 17]. T5 streamlines the model's design and makes it more adaptable to different tasks by articulating them all in the same text-to-text way. This requires only minor task-specific alterations. The detailed architecture is shown in Figure 2.

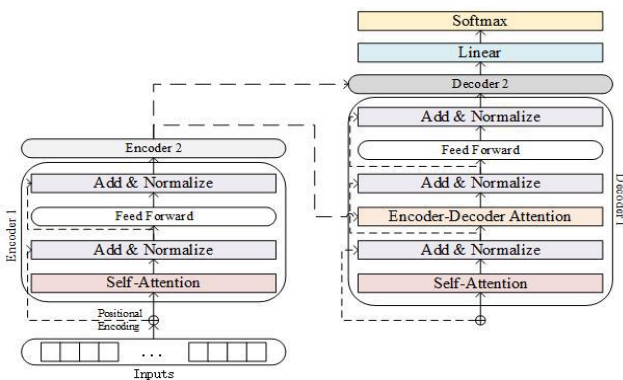


Figure 2. T5 model architecture

3.3 PEGASUS (Pre-training for abstractive summarization using extracted gap-sentences)

PEGASUS [4] represented by the model 'google/pegasus-cnn_dailymail,' is specifically created for tasks requiring abstractive text summary. During pre-training, it uses a gap-sentence generation aim, which enables it to efficiently capture the document-level structure. The unique PEGASUS objective, which involves randomly hiding off sentences in a document and training the model to produce those sentences, is where the essential innovation lies. This goal can be expressed numerically as:

$$L(\theta) = - \sum_{i=1}^N \log p(y_i | x_{masked}, y_{gold}; \theta) \quad (2)$$

where,

N represents the number of training examples.

x_{masked} : denotes the document with masked sentences.

y_{gold} : represents the target (gold) summary.

θ denotes the model parameters.

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) is a pretrained language model intended for abstractive summarization. According to its architecture, a novel gap-sentence generation technique is used to pre-train the model on a sizable corpus of documents. By inserting sentences that have been arbitrarily removed from documents, PEGASUS helps the model comprehend context and produce coherent summaries. PEGASUS is effective for automated text summarization tasks because of this method's ability to capture long-range dependencies and generate abstractive summaries that preserve the important information from the input text.

By successfully modelling the creation of short and cogent summaries, PEGASUS excels at abstractive summarization [12]. The method used in this model closely matches the fundamental characteristics of summarizing jobs.

These models, each with a distinct architecture and learning goals, make a substantial contribution to our research's automated summarization process, demonstrating the breadth and diversity of natural language processing methodologies.

4. METHODOLOGY

The suggested methodology aims to summarize business articles using the text summarization models BART, T5 and Pegasus. This study further aims to evaluate these models based on ROUGE [5] score and METEOR [6] to determine assess the limitations and strengths of each model. This procedure entails the following carefully curated steps to obtain a useful analysis as shown in Figure 3.

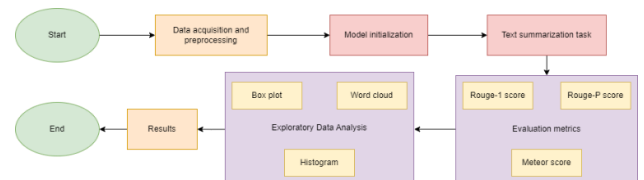


Figure 3. Steps to obtain a useful analysis

4.1 Dataset acquisition and preprocessing

This step serves as a crucial step in obtaining an accurate model and analysis. The dataset for this study is acquired from a reputable source i.e., BBC News Articles. This study is focused on the Business genre of the articles. The dataset consisted of news articles and summaries. This study aims to generate new summaries using the BART, T5 and Pegasus model and compare the generated summaries to the already existing summaries. Further, to perform this task efficiently we have used the methodical cleaning process.

4.2 Model initialization

This is the core of the methodology as it involves ensuring that the required packages are already installed and involves using these packages to import the models this study aims to compare. The required packages are transformers, rouge and sentencepiece.

4.2.1 BART model

This involves the employment of BART model and then initializing the tokenization. We initialized the model and tokenizer from the hugging face library using a pretrained BART model namely facebook/bart-large-cnn. Further we used this model to generate news article summaries.

4.2.2 T5 model

This involves importing a pretrained model from the hugging face library. The T5 model used in this study is T5small. This also loads tokenizers from the library to preprocess the input text into suitable tokens. Further this model is used to generate the appropriate news article summaries.

4.2.3 Pegasus model

Using the hugging face library, we import a pretrained Pegasus model namely google/pegasus-cnn_dailymail. This step involves the initialization of the model and tokens to be used.

4.3 Text summarization task

The input dataset is passed through the three models namely BART, T5 and Pegasus Model. The news articles and summaries are provided as an input. These models generate summaries for all the news articles inputted to the models. Each of these generated summaries are stored using indexing thereby making them easy to access and identify.

These model generated summaries are evaluated against the already existing summaries to evaluate the models performance.

4.4 Evaluation metrics

The following evaluation metrics [18] were used to assess and perform a comparison between the models.

4.4.1 Rouge-1

In the field of natural language processing, ROUGE-1, a variation of the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric [5], is an extensively used evaluation measure, particularly in the evaluation of automatic text summarization systems. Its main objective is to measure the degree of similarity and overlap between the words (or unigrams) in the automatically generated summary and those in the reference summary that was manually crafted. ROUGE-1 is specifically concerned with assessing the sufficiency of individual words in the generated summary in relation to the reference summary.

$$\text{Rouge 1} = \frac{\text{Total no. of unigrams in ref. summary}}{\text{No. of overlapping unigrams in ref and generated summary}} \quad (3)$$

4.4.2 Meteor score

METEOR [6] is an evaluation metric created to evaluate the calibre of text produced by machines (such as summaries and translations) by comparing it to a reference text, often a text produced by humans or the "gold standard." METEOR emphasises linguistic and semantic similarity, offering a more comprehensive evaluation than straightforward n-gram based metrics.

$$\text{Meteor} = (1 - \text{Penalty}) \left(\frac{\text{Precision} * \text{Recall}}{(1 - \alpha) * \text{Recall} + \alpha * \text{Precision}} \right) \quad (4)$$

where,

Precision: The ratio of the number of matched unigrams to the total number of unigrams in the generated text.

Recall: The ratio of the number of matched unigrams to the total number of unigrams in the reference text.

Penalty: The penalty term to account for unaligned words.

α : A tuneable parameter to control the importance of precision versus recall.

4.4.3 Rouge-P

In the realm of natural language processing, ROUGE-P, a ROUGE [5] metric version, is a widely used evaluation metric, particularly for evaluating the calibre of automatically generated summaries. ROUGE-P measures the overlap of n-grams (contiguous sequences of n items, often words) between the machine-generated summary and the reference (human-written) summary with a focus on the precision component of the evaluation.

ROUGE-P analyses an automated summarising system's capacity to generate n-grams that are found in the reference summary. The generated summary more closely resembles the reference summary in terms of shared n-grams the higher the ROUGE-P score.

$$\text{Rouge - p} = \frac{\text{total no. of n - grams in the generated summary}}{\text{No. of overlapping n - grams in the generated and reference summary}} \quad (5)$$

4.5 Exploratory data analysis

Figure 4 shows the creation of summaries for all articles, an exploratory data analysis (EDA) phase is carried out to learn more about the outcomes of the summary [7] and the distribution of the different parameters. The retrieved summaries serve as the starting point for collecting insights, patterns, and trends from the summarised information, along with any embedded metadata [19, 20].

Figure 4 represents the word cloud of the articles. A word cloud representation helps us understand the frequently repeated words in our dataset.

Figure 5 histogram gives us the distribution of summary lengths as we can infer from the image that most of the summaries range between 100 to 175 words.

The boxplot in Figure 6 shows us the distribution of article length and summary lengths. As inferred from the image and calculated the average article length is 329 words and average summary length is 140 words.



Figure 4. Word cloud

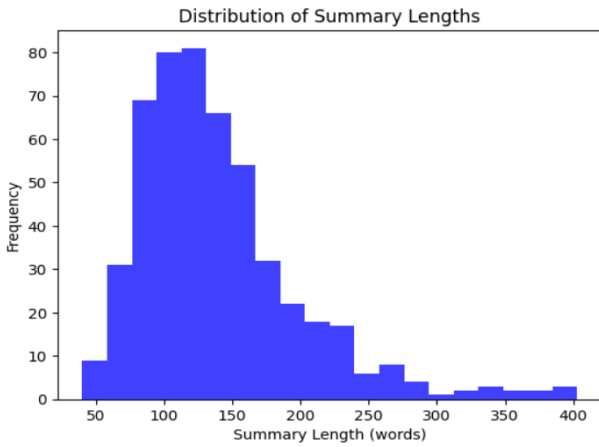


Figure 5. Histogram representing distribution of summary lengths

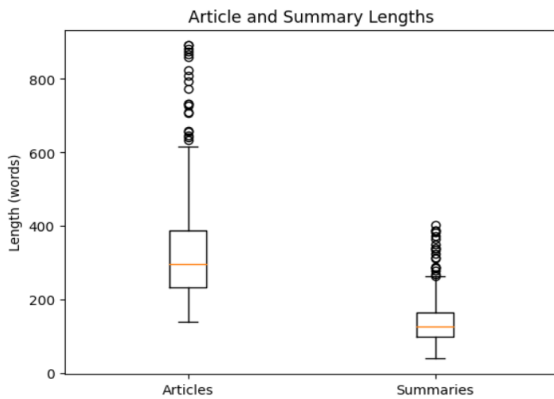


Figure 6. Boxplot representing article and summary lengths

5. RESULTS

In this paper, we carried out a thorough analysis of three cutting-edge models of abstractive summarization: BART, T5, and PEGASUS. The study, which sought to gauge these models' capacity for summarization, was based on a number of significant criteria, including METEOR, ROUGE-P, and ROUGE-1.

5.1 Meteor scores

Meteor scores often range between 0 and 1, although it isn't universally defined what is the range of a good meteor score, but it is considered that a meteor score greater than 0.25 is considered acceptable in this study. By comparing the terminology and phrasing of generated summaries to human-written reference summaries, METEOR evaluates the generated summaries' quality. With the greatest METEOR score, T5 emerged as the winner, closely followed by BART. PEGASUS has a marginally lower score, indicating that its vocabulary and phrase alignment could use some improvement as shown in Table 2.

Table 2. Comparison of average METEOR scores

METEOR Scores	
BART	0.28
T5	0.38
PEGASUS	0.25

5.2 Rouge-P scores

The generated summaries' ability to accurately extract n-grams from the reference summaries is measured by ROUGE-P (Precision). BART demonstrated the highest level of precision, demonstrating its capacity to accurately copy information from the reference summary. T5 and PEGASUS came in second and third, respectively, demonstrating a marginally lower precision, as shown in Table 3.

Table 3. Comparison of average ROUGE-P scores

ROUGE-P Scores	
BART	0.708
T5	0.685
PEGASUS	0.658

5.3 Rouge-1 scores

Rouge scores in general range from 0 to 1, (rouge-1 and rouge-p included). A rouge score that ranges between 0 to 0.3 is considered as poor, 0.3 to 1 is considered acceptable in this study. Between the generated summaries and the reference summaries, ROUGE-1 measures the unigram overlap. T5 had the greatest unigram overlap in this area BART took second place, while PEGASUS, indicating the ability to align unigrams, as shown in Table 4.

Table 4. Comparison of average ROUGE-I scores

ROUGE-I Scores	
BART	0.308
T5	0.353
PEGASUS	0.245

5.4 Compare and contrast

BART, T5, and PEGASUS have all shown considerable gains in their performance after fine-tuning for summarizing news business items. Figure 7 shows a sample summary output of the models. With the highest METEOR score, T5 distinguishes out as having a thorough comprehension of language and semantics. BART performs well in ROUGE-1, showing significant word overlap with reference summary, while T5 performs well in ROUGE-P, indicating superior precision. PEGASUS is a good option since it retains good precision despite having somewhat lower ratings in other areas. Depending on the precise requirements of the summarizing assignment, the decision between these models should be based on priorities-comprehensive understanding, precision, or a balance between the two.

We would recommend the usage of T5 model for news summarization because of its high performance in ROUGE-1 and also shows us the best results in METEOR scores as well. Though it lagged in ROUGE-P scores, the scores were good enough. As we know, METEOR scores and ROUGE-1 scores higher than 0.30 are considered good.

6. CONCLUSION

Abstractive and extractive summarization techniques have several real-life applications, such as News Summarization, Meeting Summarization, Legal Case Summarization, etc. In this study, the abstractive summarization performance of three

top models-BART, T5, and PEGASUS-was thoroughly assessed. Essential metrics METEOR, ROUGE-P (Precision), and ROUGE-1 served as the foundation for the evaluation. We evaluated them in order to determine their relative performance and to help practitioners choose the best model for their summarising requirements.

In line with human-written summaries, T5 has the highest METEOR score, which was revealed by our research. This indicates that it has the best vocabulary and phrasing alignment. BART, on the other hand, demonstrated the highest level of precision (ROUGE-P), excelling in the exact

replication of reference n-grams. A significant alignment with the reference summaries is also indicated by the fact that BART consistently produced the largest unigram overlap (ROUGE-1).

In conclusion, the particular needs of the work at hand determine the best summarization model to use. BART is a strong contender if the goal is to faithfully recreate reference content. T5 is a strong option, though, if lexical and phrase fidelity are the primary concerns. PEGASUS offers a balanced performance and may represent a workable compromise, although somewhat falling short in some criteria.

```

Article 440 Summaries:
Original Summary:
Although China's overall trade surplus is expanding, according to Chinese government figures, the Commerce Department revealed the US's deficit with China was $19.6bn in November, down from $19.7bn the month before. Against the pound, the dollar was down about 0.7% at $1,8923. But the deficit with Japan was at its worst in more than four years. The gap between US exports and imports has widened to more than $60bn (£31.7bn), an all-time record. One small bright spot for US policy-makers was a slight decline in the deficit with China, often blamed for job losses and other economic woes. By 1650 GMT, the dollar was trading against the euro at $1.3280, almost a cent and a half weaker than before the announcement. But the numbers suggested the sliding dollar - which makes exports less expensive - has had little impact, and could indicate slowing economic growth. The trade deficit is a large part of the latter.

Generated Summary (BART):
US trade deficit widens to more than $60bn (£31.7bn), an all-time record. Part of expanding deficit came from high prices for oil imports. Treasury Secretary John Snow put a brave face on the news, saying it was a sign of strong economic expansion.

Generated Summary (T5):
exports down 2.3% to $95.6bn, while imports grew 1.3% to $155.8bn. the sliding dollar - which makes exports less expensive - has had little impact. the dollar was trading against the euro at $1.3280, almost a cent and a half weaker than before.

Generated Summary (PEGASUS):
Gap between US exports and imports has widened to more than $60bn.<n>Part of expanding deficit came from high prices for oil imports.<n>The dollar was trading against the euro at $1.3280, almost a cent and a half weaker than before.

```

Figure 7. Sample summary output

7. FUTURE WORK

In the future, investigating ensemble strategies that take advantage of the capabilities of several models may lead to improvements in abstractive summarization. Additionally, examining these models across a wider range of datasets and domains may offer a more thorough knowledge of their strengths and weaknesses.

Future studies could examine real-world situations and explore how these models might be used to improve the practical implementation of these models. Valuable insights could be gained by identifying the research that is required to enable the integration of these models into practical applications. By setting the foundation for wise decision-making in abstractive summarizing, this work supports ongoing efforts to improve the caliber and effectiveness of automated summarization systems. Furthermore, evaluating the moral ramifications and any biases related to the deployment of abstractive summarization models in practical settings is an important area for further research. To guarantee responsible deployment and avoid unexpected repercussions, research addressing fairness, accountability, and transparency in automated summarization systems is crucial.

REFERENCES

- [1] Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., Huang, X. (2020). Extractive summarization as text matching. arXiv Preprint arXiv: 2004.08795. <https://doi.org/10.48550/arXiv.2004.08795>
- [2] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv Preprint arXiv: 1910.13461. <https://doi.org/10.48550/arXiv.1910.13461>
- [3] Wang, M., Xie, P., Du, Y., Hu, X. (2023). T5-Based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. Applied Sciences, 13(12): 7111. <https://doi.org/10.3390/app13127111>
- [4] Zhang, J., Zhao, Y., Saleh, M., Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning. PMLR, pp. 11328-11339. <https://doi.org/10.48550/arXiv.1912.08777>
- [5] Banerjee, S., Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of The Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72.
- [6] Lin, C.Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pp. 74-81.
- [7] Deokar, V., Shah, K. (2021). Automated text summarization of news articles. International Research Journal of Engineering and Technology, 8(9): 1-13.
- [8] Liu, Y. (2019). Fine-tune BERT for extractive summarization. arXiv Preprint arXiv: 1903.10318. <https://doi.org/10.48550/arXiv.1903.10318>
- [9] Suleiman, D., Awajan, A. (2020). Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. Mathematical Problems in Engineering, 2020: 1-29. <https://doi.org/10.1155/2020/9365340>

- [10] Xiao, W., Beltagy, I., Carenini, G., Cohan, A. (2021). PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. arXiv Preprint arXiv: 2110.08499. <https://doi.org/10.48550/arXiv.2110.08499>
- [11] Mercan, Ö.B., Cavsak, S.N., Deliahmetoglu, A., Tanberk, S. (2023). Abstractive text summarization for resumes with cutting edge NLP transformers and LSTM. arXiv Preprint arXiv: 2306.13315. <https://doi.org/10.48550/arXiv.2306.13315>
- [12] Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv Preprint arXiv: 1602.06023. <https://doi.org/10.48550/arXiv.1602.06023>
- [13] Kryściński, W., McCann, B., Xiong, C., Socher, R. (2019). Evaluating the factual consistency of abstractive text summarization. arXiv Preprint arXiv: 1910.12840. <https://doi.org/10.48550/arXiv.1910.12840>
- [14] Paulus, R., Xiong, C., Socher, R. (2017). A deep reinforced model for abstractive summarization. arXiv Preprint arXiv: 1705.04304. <https://doi.org/10.48550/arXiv.1705.04304>
- [15] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. <https://doi.org/10.48550/arXiv.1910.10683>
- [16] Dharrao, D., Bongale, A.M., Kadalaskar, V., Singh, U., Singharoy, T. (2023). Patients' medical history summarizer using NLP. In 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS), Kochi, India, pp. 1-6. <https://doi.org/10.1109/AICAPS57044.2023.10074336>
- [17] Singh, J., Patel, T., Singh, A. (2023). Performance analysis of large language models for medical text summarization. <https://doi.org/10.31219/osf.io/kn5f2>
- [18] Haque, S., Eberhart, Z., Bansal, A., McMillan, C. (2022). Semantic similarity metrics for evaluating source code summarization. In Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension, Pittsburgh, PA, USA, pp. 36-47. <https://doi.org/10.1145/3524610.3527909>
- [19] Gite, S., Patil, S., Dharrao, D., Yadav, M., Basak, S., Rajendran, A., Kotecha, K. (2023). Textual feature extraction using ant colony optimization for hate speech classification. *Big Data and Cognitive Computing*, 7(1): 45. <https://doi.org/10.3390/bdcc7010045>
- [20] Rai, S., Sharma, D. (2023). Summarization for news article using unsupervised learning techniques. In International Conference on Advances in Data-driven Computing and Intelligent Systems, pp. 169-179.