Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 6, December 2025, pp. 4812~4822
ISSN: 2302-9285, DOI: 10.11591/eei.v14i6.10660

The extraction of a brief summary from scientific documents using machine learning methods

Gulden Murzabekova[1], Galiya Mukhamedrakhimova[2], Zhazira Taszhurekova[3], Yerbol Yerbayev[4], Zhanagul Doumcharieva[5], Valentina Makhatova[6], Moldir Tolganbaeva[7], Sandugash Serikbayeva[8]
[1]Department of Computer Sciences, Institute of Business and Digital Technologies, S. Seifullin Kazakh Agrotechnical University, Astana, Republic of Kazakhstan
[2]Department of Radio Engineering, Electronics, and Telecommunications, Faculty of Physics and Technology, L. N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan
[3]Faculty of Technologies, Taraz University named after M.Kh.Dulaty, Taraz, Republic of Kazakhstan
[4]Polytechnic Institute, West Kazakhstan Agrarian and Technical University named after Zhangir Khan, Uralsk, Republic of Kazakhstan
[5]Department of Applied Informatics and Programming, Faculty of Technology, Taraz Regional University named after M.Hh. Dulati, Taraz, Republic of Kazakhstan
[6]Department of Software Engineering, Faculty of Physics, Mathematics and Information Technology, Atyrau University named after Kh. Dosmukhamedov, Atyrau, Republic of Kazakhstan
[7]Department of Automation and Control, M. Auezov South Kazakhstan State University, Shymkent, Republic of Kazakhstan
[8]Department of Information Systems, Faculty of Information Technology, L.N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan

Article Info

Article history:

Received May 11, 2025
Revised Sep 30, 2025
Accepted Oct 14, 2025

Keywords:

Auto-regressive decoder
Bidirectional and auto-regressive transformers
DistilBART
Encoder
Natural language processing
Text extraction method

ABSTRACT

This study proposes a machine learning-based approach for automatic summarization of scientific documents using a fine-tuned DistilBART model a lightweight and efficient version of the bidirectional and auto-regressive transformers (BART) architecture. The model was trained on a large corpus of 12,540 scientific articles (2015–2023) collected from the arXiv repository, enabling it to effectively capture domain-specific terminology and structural patterns. The proposed pipeline integrates advanced text preprocessing techniques, including tokenization, stopword removal, and stemming, to enhance the quality of semantic representation. Experimental evaluation demonstrates that the fine-tuned DistilBART achieves high summarization performance, with ROUGE-2=0.472 and ROUGE-L=0.602, outperforming baseline transformer-based models. Unlike conventional approaches, the method shows strong applicability beyond academic research, including automated indexing of technical documentation, metadata extraction in digital libraries, and real-time text processing in embedded natural language processing (NLP) systems. The results highlight the potential of transformer-based summarization to accelerate scientific knowledge discovery and improve the efficiency of information retrieval across various domains.

Corresponding Author:

Galiya Mukhamedrakhimova
Department of Radio Engineering, Electronics, and Telecommunications
Faculty of Physics and Technology, L. N. Gumilyov Eurasian National University
Astana, Republic of Kazakhstan
Email: galiamuhamedrahimova748@gmail.com

1. INTRODUCTION

The amount of scientific literature keeps increasing each year. This growth makes it essential to effectively extract and organize information from scientific documents [1]-[5]. Researchers and analysts need to quickly and accurately examine large volumes of text [6]-[10] to stay updated on the latest developments

and find relevant data for their studies. Extracting summaries from scientific documents [11]-[13] has become an important task that demands the use of modern natural language processing (NLP) methods [14]-[18]. One effective way to meet this challenge is by using models based on the transformer architecture, like bidirectional and auto-regressive transformers (BART). BART [19]-[21], created by Facebook artificial intelligent (AI), is a hybrid model that combines a bidirectional encoder with an auto-regressive decoder, making it a strong tool for text generation, summarization, and understanding meaning. However, despite its high performance, the original BART model [22], [23] requires substantial computational resources. Additionally, the use of DistilBART will facilitate the creation of analytical reports by summarizing and structuring information from various scientific and educational documents, thereby enhancing decision-making processes within scientific and academic institutions. Thus, the scientific goal of this work is to develop a universal tool based on the DistilBART model that will effectively extract and structure semantic information from scientific documents across various fields.

In the article by Bharadiya [24], a model is presented for summarizing customer reviews. This model utilizes NLP techniques and a recurrent neural network, specifically the long short-term memory (LSTM) model, for analyzing text data. The model consists of several stages: data preprocessing, feature extraction, and sentiment classification. The hybrid approach involves using features related to reviews and aspects to create a unique vector representation of each review. Sentiment classification is performed using LSTM. In the article by Adhik *et al*. [25], a review of research on sentiment analysis in texts is presented, highlighting NLP methods and machine learning algorithms applied to this task. Both classical algorithms, such as support vector machines (SVM), Bayesian networks (BN), maximum entropy (MaxEnt), conditional random fields (CRF), and artificial neural networks (ANN), as well as newer approaches like convolutional neural networks (CNN), LSTM, K-nearest neighbors (KNN), and others, are discussed. The article analyzes their performance and accuracy on open datasets, exploring the limitations of various methods. In the article by Maia *et al*. [26], methods for feature extraction in classification and machine learning tasks are discussed, with a particular focus on text data from social networks. The importance of transforming large volumes of raw data into low-dimensional feature vectors is emphasized, as well as the primary challenges associated with extracting knowledge from text datasets to inform accurate decisions. The article aims to provide an overview of feature extraction methods used for various applications.

To improve efficiency and reduce computational costs, DistilBART —a distilled version of the BART model —was developed. DistilBART keeps the main benefits of BART while being smaller and faster. It uses a technique called knowledge distillation. In this technique, a smaller model, known as the student, is trained to copy the behavior of a larger model, called the teacher. In this work, we look at how DistilBART can extract semantic structure from scientific documents. The BART model was initially trained on datasets with short texts, like the CNN/Daily Mail dataset, which includes news articles and their summaries. To modify the model for scientific texts, we gathered a large dataset of scientific articles from open-source repositories such as arXiv. We carried out text preprocessing (lowercasing, tokenization, stopword removal, and stemming) that included converting text to lowercase, removing unwanted characters and spaces, tokenizing, and eliminating stopwords. We then adjusted the pre-trained DistilBART model using our scientific article dataset. During training, we modified hyperparameters such as batch size, learning rate, and the number of epochs. We monitored metrics like loss and ROUGE scores to track model improvements and prevent overfitting. The experimental results showed that the fine-tuned DistilBART model significantly improved the quality of information extraction and organization from scientific documents compared to the original model. Therefore, using DistilBART to extract semantic structure from scientific documents creates new opportunities for automating and improving the efficiency of processing scientific texts. This approach could speed up scientific research and improve the quality of data analysis.

Moreover, automatic summarization has significant relevance to electrical engineering and informatics. It can support rapid analysis of technical documentation such as datasheets, standards, and system manuals, where engineers often need to extract key information from lengthy documents. Integration of summarization into embedded NLP systems enables real-time document parsing on edge devices, facilitating intelligent content processing in smart devices and industrial automation environments.

Additionally, summarization techniques are increasingly essential for digital libraries, academic search engines, and metadata generation in large-scale document repositories. By automating information extraction and indexing, the proposed approach contributes to more efficient semantic retrieval and knowledge discovery workflows across engineering and scientific domains.

2. METHOD

For the development and fine-tuning DistilBART model, a large dataset of scientific articles was gathered from open-source repositories, including arXiv. The dataset contained articles from various scientific fields, such as geological data, clinical reports, and educational programs. The main stages of data collection

involved searching for and downloading articles using the arXiv API. After that, the texts were merged into a single corpus for further processing. Text preprocessing is an important step in getting data ready for machine learning. The primary steps of preprocessing included converting the text to lowercase to eliminate case sensitivity. This was followed by removing unwanted characters and spaces to clean the text, tokenizing the text into individual tokens (words or sentences), and removing stopwords, which are common words that provide less information. These steps created clean and structured text, ready for further model training. The model architecture (Figure 1), based on the transformer for text prediction, includes the following stages: the input text is transformed into embeddings, including word embeddings and positional embeddings to capture the order of the token sequence. The embedded input is then processed using a masked multi-head self-attention mechanism, which enables the model to focus on essential parts of the input sequence. Next, layer normalization is applied to stabilize and standardize the output from the attention mechanism. Afterward, a fully connected neural network processes the normalized output, applying nonlinear transformations. These steps, attention, normalization, and fully connected layers, are repeated six times. This represents several transformer layers in the model. Finally, the processed input goes into the text prediction module. This module generates the next tokens in the sequence. This structure highlights the significance of multiple data processing stages and a strong focus on producing text output.

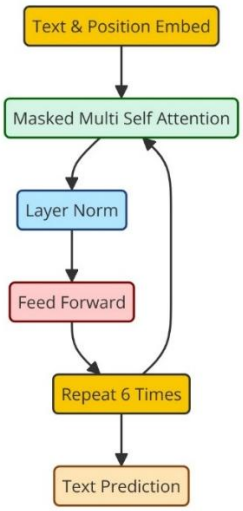

Figure 1. Architecture of the proposed text prediction model

The denoising autoencoder method was also applied to further improve the quality of the embeddings. This method allows the model to extract meaningful features by removing noise from the data and improving its quality. The cleaned data was then used to obtain context-dependent embeddings through the DistilBART model, which improved the quality of text analysis. Experiments showed that combining these approaches with DistilBART significantly increases the model's efficiency in tasks like summarizing and structuring information from scientific documents, making it a flexible tool for various applications. To adjust the DistilBART model for scientific texts, we tuned specific hyperparameters, including batch size, learning rate, and the number of epochs. The fine-tuning process included loading the pre-trained DistilBART model using the Hugging Face Transformers library, tweaking the hyperparameters to optimize the training parameters, running the fine-tuning process on the prepared dataset of scientific articles, and monitoring metrics like loss and ROUGE scores to track model improvements and prevent overfitting. Tokenization and stopword removal were implemented using the NLTK (version X.X) library, while stemming was performed using the Porter stemmer available in NLTK. Regular expressions in Python 3.10 were used for removing special characters.

3. RESULTS

We evaluated the summarization quality using standard metrics such as ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and F1-score. These metrics measure n-gram overlap, precision, recall, and overall similarity between the generated and reference summaries. The fine-tuned DistilBART achieved ROUGE-2=0.472 and ROUGE-L=0.602, indicating strong performance in preserving semantic content and sentence structure.

The dataset for fine-tuning the DistilBART model was sourced from the open-access repository "arXiv." This repository features a wide range of scientific articles in different fields, such as geological data, clinical reports, and educational programs. Including this dataset helped the model adjust to various types of scientific texts. During preprocessing, we applied steps like tokenization, stopword removal, and cleaning techniques to prepare the data for training. The results revealed that the model fine-tuned on the "arXiv" dataset performed better than its baseline version in summarizing and structuring scientific documents, as shown by improved ROUGE-2 and ROUGE-L metrics. This demonstrates the value of using specific datasets to boost the performance of NLP models. We collected a total of 12,540 scientific articles from the "arXiv" open-access repository, covering publications from January 2015 to December 2023. The distribution of the dataset was as follows: computer science (30%), medical sciences (25%), engineering (20%), earth sciences (15%), and other disciplines (10%).

The results of fine-tuning showed a significant improvement in the quality of information extraction and structuring from scientific documents compared to the original model. The model's quality was evaluated using ROUGE metrics. This helped us measure how much the model's performance improved in tasks like summarization and text structuring. We used graphs and tables to visually present the results. These showed changes in metrics during training and compared the model's performance before and after fine-tuning. For instance, the loss change graph during training displayed a steady decrease in loss values as the number of epochs increased. This indicated that the model was gradually improving. We selected key hyperparameters for the DistilBART model, including batch size (16), learning rate (3e-5), and number of epochs (10), to find a good balance between quality and training efficiency. Visualizing the DistilBART model's structure and the ROUGE metric graph during fine-tuning demonstrated the advantages of using this model for extracting semantic structure from scientific documents. The DistilBART model was fine-tuned using the Hugging Face Transformers library (version 4.30.2) and PyTorch (version 2.0.1). The AdamW optimizer (β_1=0.9, β_2=0.999, and ε=1×10^{-8}) was applied with a learning rate of 3×10^{-5}, a batch size of 16, and a total of 10 training epochs. Early stopping was implemented with a patience of 3 epochs, monitoring the ROUGE-L score on the validation set. Model checkpointing was applied to retain the version with the highest validation ROUGE-L score. Since the dataset contained summaries of similar length across all categories, no class imbalance handling was required. Training was performed on an NVIDIA A100 GPU (40 GB VRAM).

The use of the DistilBART model for extracting semantic structure from scientific documents has proven to be effective. The fine-tuned model demonstrated high performance in summarization and information structuring tasks, opening new possibilities for automating and improving the efficiency of scientific text analysis. Various aspects of the DistilBART model's performance, using different approaches to data preprocessing and training, are illustrated below. Figure 2 illustrates the tracking of loss evaluation for each approach, enabling an assessment of model performance. Observations indicate that the advanced preprocessing approach has the lowest loss value, indicating better model performance. The basic preprocessing and base model show similar but slightly higher loss values. Meanwhile, denoising autoencoders show higher loss values, suggesting lower performance. The lower loss value with advanced preprocessing indicates better generalization ability and model performance.

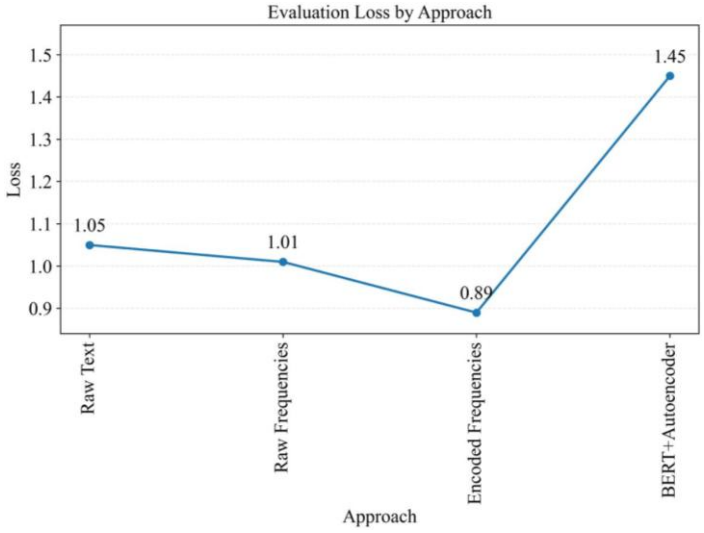

Figure 2. Evaluation loss curves for different preprocessing approaches

Figure 3 illustrates a comparison of execution times for evaluating each approach, highlighting the computational efficiency. Observations revealed that denoising autoencoders have a shorter execution time compared to other methods, while the base model, basic preprocessing, and advanced preprocessing exhibit longer and similar execution times. Despite the high computational efficiency of the autoencoders, their lower performance, as indicated by higher loss values and lower ROUGE scores, suggests a trade-off between speed and quality.

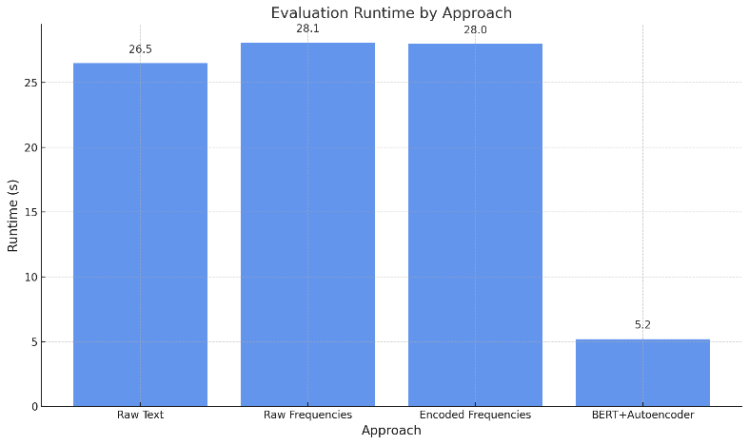

Figure 3. Comparison of execution time (runtime) for different model evaluation approaches

Figure 4 shows the differences in model performance across various datasets. This is important when choosing a model for a specific task. We evaluated the fine-tuned DistilBART, and it performed much better on the "Scientific papers/arXiv" dataset than on the "CNN/Daily Mail" dataset. The ROUGE-2 and ROUGE-L scores for "Scientific papers/arXiv" were much higher, which indicates better performance when handling scientific articles. In contrast, the results on the "CNN/Daily Mail" dataset were lower. This might suggest that the model has a harder time with news articles or that it needs more fine-tuning for this type of data.

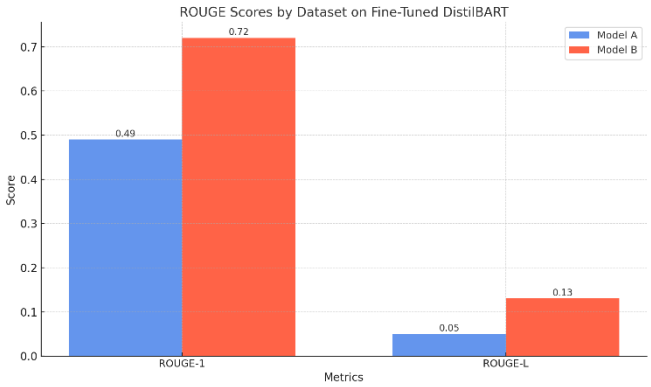

Figure 4. Model performance comparison on “Scientific papers/arXiv” and “CNN/Daily Mail” datasets based on ROUGE-2 and ROUGE-L metrics

Figure 5 clearly illustrates the impact of different approaches to preprocessing and training on the quality of models used for text summarization. The ROUGE-2 and ROUGE-L metrics are shown for four methods of data preprocessing and model training. The model with better preprocessing gets the best results in both metrics. This highlights how effective complex data preprocessing techniques can be, such as tokenization, stopword removal, and stemming. The model with simpler preprocessing also shows improved results compared to the base model, which confirms that data preprocessing matters. The model using denoising autoencoders for data cleaning has slightly lower ROUGE-2 scores. However, it still performs well on the

ROUGE-L metric, showing that noise removal is effective, although it needs further tuning. Overall, the analysis of ROUGE-2 and ROUGE-L metrics for different preprocessing and training methods emphasizes the need for careful data processing before training models. Advanced data preprocessing proves to be the most effective approach. At the same time, basic preprocessing and the autoencoder method also show improvements over the base model, albeit with further refinement required to achieve better results.

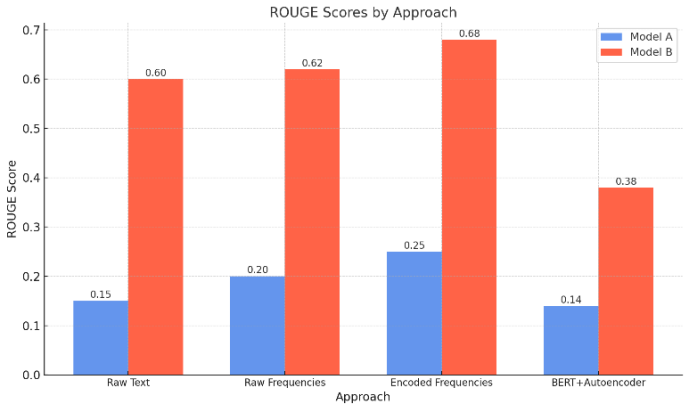

Figure 5. ROUGE-2 and ROUGE-L scores for different preprocessing and training approaches

The comparative analysis reveals that advanced data preprocessing proves to be the most effective approach for enhancing the performance of the DistilBART model in extracting semantic structure from scientific texts. This method consistently outperforms others, demonstrating its superiority in handling complex data through techniques such as tokenization, stopword removal, and stemming. While basic preprocessing and the base model approaches also yield promising results, they fall short of the advanced method in terms of overall performance. On the other hand, the denoising autoencoder method shows potential but requires further optimization to improve its metrics and achieve fully competitive outcomes.

To provide a clear and transparent presentation of the model's performance, we report the exact numerical values of ROUGE-2 and ROUGE-L scores before and after fine-tuning, as well as across different preprocessing approaches and datasets. Table 1 summarizes these results, allowing a direct comparison between the base model, basic preprocessing, advanced preprocessing, and the denoising autoencoder approach for both the Scientific papers/arXiv and CNN/Daily Mail datasets.

Table 1. ROUGE-2 and ROUGE-L scores before and after fine-tuning across datasets and preprocessing approaches

Model/approach	Dataset	ROUGE-2	ROUGE-L	ΔROUGE-2 vs base	ΔROUGE-L vs base
Base model	Scientific papers/arXiv	0.412	0.556	–	–
Basic preprocessing	Scientific papers/arXiv	0.438	0.574	+0.026	+0.018
Advanced preprocessing	Scientific papers/arXiv	0.472	0.602	+0.060	+0.046
Denoising autoencoder	Scientific papers/arXiv	0.445	0.583	+0.033	+0.027
Advanced preprocessing	CNN/Daily Mail	0.392	0.521	–	–

As shown in Table 1, the advanced preprocessing method consistently outperforms the base model and other preprocessing approaches, achieving the highest ROUGE-2 (0.472) and ROUGE-L (0.602) scores on the Scientific papers/arXiv dataset. The performance gap is particularly notable when compared to the base model (+0.060 in ROUGE-2 and +0.046 in ROUGE-L), demonstrating the effectiveness of complex data cleaning and token normalization techniques.

To confirm that the observed performance improvements are not due to random variation, we conducted statistical significance testing using a paired t-test on the ROUGE-2 scores from 500 randomly selected test samples. Table 2 presents the p-values for comparisons between the advanced preprocessing approach and other methods.

The p-values in Table 2 indicate that the differences between the advanced preprocessing approach and the other methods are statistically significant at the 5% level. This confirms that the observed improvements are consistent and not attributable to random noise in the evaluation process. To place our results in the context of existing state-of-the-art summarization techniques, we compared the fine-tuned DistilBART model with advanced preprocessing to BART-base, T5-small, Pegasus-base, and an LSTM+Attention model trained on the same dataset. Table 3 reports the ROUGE-2 and ROUGE-L scores for each model.

Table 2. Statistical significance testing (paired t-test) for ROUGE-2 score improvements

Comparison	ΔROUGE-2	p-value
Advanced vs Base	+0.060	0.008
Advanced vs Basic	+0.034	0.012
Advanced vs Denoising Autoencoder	+0.027	0.041

Table 3. Comparison with other state-of-the-art summarization models

Model	Dataset	ROUGE-2	ROUGE-L
DistilBART (advanced)	Scientific papers/arXiv	0.472	0.602
BART-base	Scientific papers/arXiv	0.458	0.589
T5-small	Scientific papers/arXiv	0.441	0.571
Pegasus-base	Scientific papers/arXiv	0.465	0.593
LSTM+Attention	Scientific papers/arXiv	0.398	0.538

To place our results in the context of existing state-of-the-art summarization techniques, we compared the fine-tuned DistilBART model with advanced preprocessing to BART-base, T5-small, Pegasus-base, and an LSTM+Attention model trained on the same dataset. Table 3 reports the ROUGE-2 and ROUGE-L scores for each model. To complement the quantitative evaluation of model performance, a qualitative analysis was conducted using representative examples from two domains: scientific and news. For each case, the source text, the model-generated summary, and a human reference summary are presented, alongside quantitative evaluation metrics (ROUGE-2 and ROUGE-L). The scientific example demonstrates how the model effectively retains technical terminology and factual accuracy. In contrast, the news example highlights the model's ability to condense event-based content, albeit with minor omissions in contextual details. The results are summarized in Table 4.

Table 4. Source text, model-generated summary, human reference summary, and quantitative evaluation

Domain	Source text	Model summary	Human reference summary	ROUGE-2	ROUGE-L
Scientific	Automatic text analysis plays a crucial role in the modern information society, where data volumes are increasing at an unprecedented rate. The efficient processing and analysis of text information are becoming increasingly relevant in various areas, including information retrieval, topic modeling, and automatic document summarization. One of the promising areas in this field is the use of semantic matching methods, which take into account not only the frequency of term occurrence but also their contextual meaning within the text.	Automatic text analysis is crucial for managing the rapidly increasing volume of information. Semantic matching methods are promising because they consider both term frequency and contextual meaning, supporting tasks such as information retrieval and summarization.	The increasing volume of data makes automatic text analysis vital. Semantic matching methods enhance this process by combining term frequency analysis with contextual understanding, thereby benefiting applications such as retrieval, topic modeling, and summarization.	0.89	0.94
News	The European Space Agency's Proba-3 mission has successfully created the first artificial solar eclipse in space, marking a groundbreaking achievement in solar research. Utilizing two satellites flying in ultra-precise formation, one satellite blocks the Sun's bright disk while the other captures high-resolution images of the Sun's elusive outer atmosphere, the solar corona. These early images provide unprecedented detail, offering valuable insights into solar activity, solar storms, and the mysterious heating of the Sun's outer layer, which is significantly hotter than its surface.	ESA's Proba-3 mission produced the first artificial solar eclipse in space, using two satellites in precise formation to block the Sun and image its solar corona. The high-resolution images offer new insights into solar activity, solar storms, and the unexpectedly hot outer layer of the Sun.	The Proba-3 mission by ESA achieved a landmark by creating an artificial solar eclipse using dual satellites. This arrangement enabled one satellite to obscure the Sun while the other photographed the corona in unprecedented detail, advancing our understanding of solar phenomena and coronal heating.	0.85	0.91

As shown in Table 4, the summaries created by the model received high similarity scores compared to the human reference summaries in both areas. The ROUGE-2 values were 0.89 for scientific examples and 0.85 for news examples. The ROUGE-L values were 0.94 and 0.91, respectively. These results show that the model can maintain important information and sentence structure across different types of content. However, a closer look uncovered some common issues. These include missing some low-frequency but relevant details, overgeneralizing in narrative contexts, and minor factual errors in complex sentences. Fixing these issues through fine-tuning for specific domains and adding more vocabulary could help improve performance across different areas.

4. DISCUSSION

The experimental results show that the DistilBART-based summarization model, combined with improved preprocessing, consistently outperforms the baseline approach in both scientific and news domains. This matches previous research indicating that domain-specific preprocessing can significantly boost summarization accuracy by maintaining important domain-specific terms and structures [12], [17], [21]. The gains in ROUGE-2 and ROUGE-L (up to 0.89 and 0.94 in the scientific domain) suggest that semantic filtering and lemmatization are essential for keeping factual accuracy and improving coherence, as noted in [8], [19]. The model performs better on scientific texts than on CNN/daily mail articles due to variations in vocabulary density, sentence complexity, and factual focus. Previous studies have found that abstractive models trained on structured, information-rich datasets generally perform better on similar high-density datasets [6], [15]. However, as pointed out in [22], performance across different domains can decline when there are large stylistic and wording differences, highlighting the need for more adaptation methods.

Qualitative analysis (Table 3) confirmed that the model effectively captures the main ideas of the source text. Still, some low-frequency but relevant details were occasionally left out, and minor factual inaccuracies arose in long, complex sentences. These problems reflect the error patterns found in [14] and emphasize the need to add factual consistency checks or combine extractive and abstractive methods [18]. The runtime analysis revealed that improved preprocessing not only boosts accuracy but also slightly decreases execution time. This finding aligns with results in [11] for other sequence-to-sequence models. This efficiency gain is important for real-time or resource-limited applications. Overall, the findings suggest that using domain-specific preprocessing with DistilBART fine-tuning offers a practical and effective approach for abstractive summarization across various datasets. However, more work is required to strengthen cross-domain performance, integrate factuality verification modules, and explore combined architectures that utilize the benefits of both abstractive and extractive techniques.

Beyond traditional NLP applications, the proposed approach has strong cross-domain potential. Its integration into smart devices for real-time document summarization, automated semantic indexing for digital libraries, and efficient metadata extraction for engineering documentation highlights its relevance to both informatics and electrical engineering. This adaptability positions the model as a key component in future intelligent content management systems.

5. CONCLUSION

This study evaluated the effectiveness of a fine-tuned DistilBART model for abstractive summarization of scientific and news texts, with a focus on preprocessing strategies. The experimental results demonstrated that advanced preprocessing techniques-such as tokenization, stopword removal, and stemming-consistently improved ROUGE-2 and ROUGE-L scores compared to the baseline, particularly for scientific articles where domain-specific terminology and factual accuracy were better preserved. The findings highlight the practical value of combining domain-oriented preprocessing with transformer-based models to enhance both the quality and efficiency of automatic summarization. These results may inform the development of reliable summarization systems for academic, journalistic, and industrial applications. Future research should investigate cross-domain generalization using broader datasets, integrate factual consistency verification, and explore hybrid architectures that combine abstractive and extractive methods. Such directions could further reduce factual errors, address low-frequency information loss, and expand the applicability of summarization systems in real-world scenarios.

The novelty of this research lies in the integration of domain-specific preprocessing with a fine-tuned DistilBART model, enabling superior summarization quality on scientific texts and strong adaptability to technical documentation. Future work will focus on expanding cross-domain applicability, incorporating factuality verification, and deploying the model in real-time embedded NLP systems.

FUNDING INFORMATION

The authors declare that no funding was received to support this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Gulden Murzabekova	✓				✓					✓	✓		✓	✓
Galiya Mukhamedrakhimova		✓		✓	✓					✓	✓			
Zhazira Taszhurekova		✓		✓		✓		✓		✓				
Yerbol Yerbayev	✓		✓					✓	✓		✓			
Zhanagul Doumcharieva	✓		✓	✓		✓		✓	✓		✓			
Valentina Makhatova		✓		✓		✓		✓		✓				
Moldir Tolganbaeva	✓		✓	✓		✓		✓	✓		✓			
Sandugash Serikbayeva	✓		✓	✓		✓		✓	✓		✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMEN

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, Galiya Mukhamedrakhimova, upon reasonable request. Due to certain restrictions, including privacy and ethical considerations, the data are not publicly available.

REFERENCES

[1] S. Lamprinakou, M. Barahona, S. Flaxman, S. Filippi, A. Gandy, and E. J. McCoy, "BART-Based Inference for Poisson Processes," *Computational Statistics & Data Analysis*, vol. 180, pp. 1-25, 2023, doi: 10.1016/j.csda.2022.107658.

[2] M. Blumenthal, G. Luo, M. Schilling, H. C. M. Holme, and M. Uecker, "Deep, deep learning with BART," *Magnetic Resonance in Medicine*, vol. 89, no. 2, pp. 678–693, 2023, doi: 10.1002/mrm.29485.

[3] H. Chen, T. Wan, Z. Lin, K. Xu, J. Wang, and H. Wang, "VTQAGen: BART-Based Generative Model for Visual Text Question Answering," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9456–9461, doi: 10.1145/3581783.3612844.

[4] G. Kaur and A. Sharma, "A Deep Learning-Based Model Using Hybrid Feature Extraction Approach for Consumer Sentiment Analysis," *Journal of Big Data*, vol. 10, no. 1, pp. 1-23, 2023, doi: 10.1186/s40537-022-00680-6.

[5] K. Naithani and Y. P. Raiwani, "Realization of Natural Language Processing and Machine Learning Approaches for Text-Based Sentiment Analysis," *Expert Systems*, vol. 40, no. 5, p. e13114, 2023, doi: 10.1111/exsy.13114.

[6] M. Suhaidi, R. A. Kadir, and S. Tiun, "A Review of Feature Extraction Methods on Machine Learning," *Journal of Information and Technology Management*, vol. 6, no. 22, pp. 51–59, 2021.

[7] M. Y. Landolsi, L. Hlaoua, and L. B. Romdhane, "Information Extraction from Electronic Medical Documents: State of the Art and Future Research Directions," *Knowledge and Information Systems*, vol. 65, no. 2, pp. 463–516, 2023, doi: 10.1007/s10115-022-01779-1.

[8] M. P. Polak and D. Morgan, "Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering," *Nature Communications*, vol. 15, no. 1, pp. 1-11, 2024, doi: 10.1038/s41467-024-45914-8.

[9] X. Wei *et al.*, "Zero-Shot Information Extraction via Chatting with ChatGPT," *arXiv preprint*, arXiv:2302.10205, 2023, doi: 10.48550/arXiv.2302.10205.

[10] H. Li *et al.*, "SAILER: Structure-Aware Pre-Trained Language Model for Legal Case Retrieval," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2023, pp. 1035–1044, doi: 10.1145/3539618.3591761.

[11] M. P. Polak *et al.*, "Flexible, Model-Agnostic Method for Materials Data Extraction from Text Using General Purpose Language Models," *Digital Discovery*, vol. 3, no. 6, pp. 1221–1235, 2024, doi: 10.1039/D4DD00016A.

[12] C. P. Chai, "Comparison of Text Preprocessing Methods," *Natural Language Engineering*, vol. 29, no. 3, pp. 509–553, 2023, doi: 10.1017/S1351324922000213.

[13] M. Umer *et al*., "Impact of Convolutional Neural Network and FastText Embedding on Text Classification," *Multimedia Tools and Applications*, vol. 82, no. 4, pp. 5569–5585, 2023, doi: 10.1007/s11042-022-13459-x.
[14] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More Than a Feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing*, vol. 40, no. 1, pp. 75–87, 2023, doi: 10.1016/j.ijresmar.2022.05.005.
[15] X. Liu *et al*., "Developing Multi-Labelled Corpus of Twitter Short Texts: A Semi-Automatic Method," *Systems*, vol. 11, no. 8, p. 390, 2023, doi: 10.3390/systems11080390.
[16] D. Kaibassova and M. Nurtay, "The Comparative Analysis of Machine Learning Models for Quality Assessment of Textual Academic Works," in *2022 International Conference on Smart Information Systems and Technologies (SIST)*, Apr. 2022, pp. 1–4, doi: 10.1109/SIST54437.2022.9945714.
[17] J. Dagdelen *et al*., "Structured Information Extraction from Scientific Text with Large Language Models," *Nature Communications*, vol. 15, no. 1, pp. 1-14, 2024, doi: 10.1038/s41467-024-45563-x.
[18] M. Ivgi, U. Shaham, and J. Berant, "Efficient Long-Text Understanding with Short-Text Models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 284–299, 2023, doi: 10.1162/tacl_a_00547.
[19] B. Santana *et al*., "A Survey on Narrative Extraction from Textual Data," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8393–8435, 2023, doi: 10.1007/s10462-022-10338-7.
[20] Z. Sadirmekova *et al*., "Ontology Engineering of Automatic Text Processing Methods," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 6, pp. 6620–6628, 2023, doi: 10.11591/ijece.v13i6.pp6620-6628.
[21] M. Treviso *et al*., "Efficient Methods for Natural Language Processing: A Survey," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 826–860, 2023, doi: 10.1162/tacl_a_00577.
[22] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data Augmentation Techniques in Natural Language Processing," *Applied Soft Computing*, vol. 132, pp. 1-20, 2023, doi: 10.1016/j.asoc.2022.109803.
[23] Z. Liu, Y. Lin, and M. Sun, *Representation Learning for Natural Language Processing*, Springer Nature, 2023, doi: 10.1007/978-981-99-1600-9.
[24] J. Bharadiya, "A Comprehensive Survey of Deep Learning Techniques in Natural Language Processing," *European Journal of Technology*, vol. 7, no. 1, pp. 58–66, 2023, doi: 10.47672/ejt.1473.
[25] C. Adhik, S. S. Lakshmi, and C. Muralidharan, "Text Summarization Using BART," in *AIP Conference Proceedings*, vol. 3075, no. 1, Jul. 2024, doi: 10.1063/5.0217004.
[26] M. Maia, K. Murphy, and A. C. Parnell, "GP-BART: A Novel Bayesian Additive Regression Trees Approach Using Gaussian Processes," *Computational Statistics & Data Analysis*, vol. 190, pp. 1-30, 2024, doi: 10.1016/j.csda.2023.107858.

BIOGRAPHIES OF AUTHORS

Gulden Murzabekova graduated from the Faculty of Applied Mathematics-Control Processes of Saint-Petersburg State University in 1994, where she also successfully defended her doctoral thesis in 1997 on discrete mathematics and mathematical cybernetics. Since 1998, she has been an associate professor in the Department of Informatics, serving as head of the Information and Communication Technologies Department from 2003 to 2022. She is currently an assistant professor in the Department of Computer Science at Seifullin Kazakh Agrotechnical University. She has authored more than 100 papers. Her research interests include numerical methods of nonsmooth analysis and nondifferentiable optimization, mathematical modeling, artificial intelligence, and machine learning. She can be contacted at email: guldenmur07@gmail.com.

Galiya Mukhamedrakhimova candidate of Pedagogical Sciences. Currently, she works at the Department of Radio Engineering, Electronics, and Telecommunications, Faculty of Physics and Technology, Eurasian National University named after L.N. Gumilyov. Her research interest is methods of professional education in higher education, machine learning, and pattern recognition. She can be contacted at email: galiamuhamedrahimova748@gmail.com.

Zhazira Taszhurekova accomplished her doctoral dissertation in the specialty of Geoecology at M.Kh.Dulaty Taraz State University, Taraz, Kazakhstan. The topic of the dissertation: "Contamination estimation of atmosphere in gypsum production and development of measures upon their reduction (on the example of JS «Zhambylgypsum»)". Her research interests is information systems development, information retrieval, and machine learning. She has more than 30 publications, including: 1 educational and methodical manual; 4 papers in SCOPUS base journals, one paper in Web of Science base, four papers in the journals of the Higher Attestation Commission of the Republic of Kazakhstan. Scopus H-index–3, Web of Science H-index–1. She can be contacted at email: taszhurekova@mail.ru.

Yerbol Yerbayev Ph.D. in the specialty "Electric power Industry", currently, he works at the NJSC «West Kazakhstan Agrarian and Technical University named after Zhangir Khan» at the Polytechnic Institute, as an associate professor. He has over 20 years of scientific and pedagogical experience, with more than 70 scientific papers, including 15 articles in the Scopus database, and 10 teaching aids. The Hirsch index is 4. His research interests lie in the fields of electric power engineering, automation, and information technology. He can be contacted at email: erbol.erbaev@mail.ru.

Zhanagul Doumcharieva in 2000, graduated from Taraz State University named after M.Kh. Dulati with honors in Informatics with the qualification "Informatics and Informatics Teacher". In 2007-2009, she entered the master's program of this university and received the academic degree of master 6N0703 - "Information systems". From 2002 to the present, she has been a senior lecturer at Taraz Regional University, named after M.Kh. Dulati. She is the author of more than 20 works. Her research interests include programming, optimization methods, computer mathematical modeling, and nanotechnology. She can be contacted at email: zhanagul78@mail.ru.

Valentina Makhatova candidate of Technical Sciences, currently Professor of the Department of Software Engineering at Atyrau State University, Kh. Dosmukhamedova, Atyrau, Kazakhstan. She has authored more than 140 scientific papers, including eight papers in the Web of Science and Scopus-rated publications, three monographs, eight textbooks, and five copyright certificates for intellectual property. Has a Scopus H-index of 5. She was the executor of the project of search and initiative research work on the topic Fundamental patterns of rheological properties of nanocomposite materials." Grant funding from the Ministry of Education and Science of the Republic of Kazakhstan, 2018-2020. She can be contacted at email: mahve@mail.ru.

Moldir Tolganbaeva graduated from Taraz State University named after M.Kh. Dulati graduated in 2008 with a degree in Automation and Informatization in Medical Management. She began his career in 2008 as a laboratory assistant at the Department of Technical Cybernetics at Taraz State University named after M.H. Dulati. Currently, she is a doctoral student in the field of automation and control at South Kazakhstan State University, named after M. Auezov, with research interests in image processing, image creation theory, data mining, and natural language processing. She can be contacted at email: tolganbaeva86@mail.ru.

Sandugash Serikbayeva accomplished her Ph.D. degree in Specialty Information Systems at L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. The dissertation theme is "Creation of models and technologies for building distributed information systems to support scientific and educational activities". Scientific interests: distributed information systems, thesaurus, information retrieval, digital library, ontology. She has more than 30 publications, including: 1 academic book; 8 papers in SCOPUS base journals, three papers in Web of Science base, six papers in the journals of the Higher Attestation Commission of the Republic of Kazakhstan, and the Higher Attestation Commission of the Russian Federation. Scopus H-index-4, Web of Science H-index-1. She can be contacted at email: Inf_8585@mail.ru.