WHICH AI MODEL LEADS IN SUMMARIZING FINANCIAL ARTICLES? A COMPARATIVE ANALYSIS OF GPT, MISTRAL, AND LLAMA

Jana Dannhoferová[1], Jan Přichystal[1]

[1]Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

ABSTRACT

In an era where financial data grows at an unprecedented pace, effective summarization is vital for informed decision-making. This study rigorously evaluates the summarization capabilities of three advanced AI models—GPT-4o, Mistral Instruct, and Llama 3.1 8B Instruct 128k—when applied to diverse financial articles. A key contribution of this research is the development of a comprehensive evaluation framework, which assesses the models across critical dimensions, including accuracy, clarity, relevance, adherence to formatting specifications, and practical usability. While GPT consistently achieved the highest overall scores, Llama demonstrated superior performance in certain criteria, such as clarity and compression efficiency, highlighting its potential for applications where brevity and conciseness are prioritized. Despite occasional inconsistencies, Mistral excelled in generating concise summaries with high compression ratios. Our findings emphasize that the selection of an AI model should depend on specific task priorities—whether it is accuracy, brevity, or response speed. These insights underline the importance of both rigorous evaluation methodologies and careful model selection based on task-specific requirements, paving the way for more targeted applications and further research into AI-driven summarization tools in finance.

Keywords: artificial intelligence, LLMs, AI summarization, financial articles, evaluation, GPT, Mistral, Llama, metrics

JEL Code: O33, J24

1 INTRODUCTION

Summarizing long pieces of text is a principal task in natural language processing with Machine Learning-based text generation models such as Large Language Models (LLM) being particularly suited to it (Dhaini *et al.*, 2024). The process of text summarization is one of the applications of natural language processing (NLP) that presents one of the most challenging obstacles (Saiyyad & Patil, 2023). Current research in NLP has mainly focused on the general capabilities

https://doi.org/10.11118/978-80-7701-047-4-0039

of language models in text summarization, but much less attention has been paid to the application of these models to summarization of specific articles that have specific characteristics. For example, financial articles often contain technical terms, numbers, abbreviations, graphs, and statistics that require deeper contextual interpretation. Specific content, such as descriptions of market performance or risk analysis, can be difficult for traditional summarization models that are not trained on domain data. Finally, often the summary needs to contain not only the key points, but also some subtle meanings or conclusions. Research such as benchmark tests of language models (e.g. GPT, LLaMA, Mistral) typically focus on general datasets (CNN/DailyMail, XSum) that do not contain expert articles. It is unclear how these models perform when summarizing financial text, and whether they need further adaptation (fine-tuning) to perform better in this domain.

Despite significant advancements, text summarization continues to encounter various challenges and limitations. Persistent issues include the potential loss of essential information, semantic inconsistencies in longer summaries, and the need for domain-specific knowledge (Supriyono *et al*., 2024). Many companies and institutions work with sensitive or confidential data that must be protected and cannot be shared with external service providers using global language models. These organizations are therefore faced with the need for local solutions that guarantee a higher level of data security. Local models such as Mistral are an effective alternative because they can be deployed directly on secure internal servers and eliminate the risk of data leakage. This research aims to verify whether these local solutions can compete with the performance of global models in the field of abstract summarization of expert texts.

Generative AI models, such as neural networks and deep learning architectures, are employed to extract salient information and generate coherent summaries that capture the essence of the original articles (Roy *et al*., 2023). Different language models are trained on different datasets and with different architectures, which affects their capabilities and specialization. For example, some models focus on general text comprehension, while others are better suited for analytical tasks or for processing specific types of text. This diversity means that each model is better suited to different types of tasks, including different approaches to text summarization.

The aim of this paper is to analyze and compare the capabilities of selected language models (Mistral, LLaMA and GPT) in summarizing financial expert articles according to the designed methodology.. The research focuses on:

- Development of a unique summarization quality assessment methodology.
- Evaluating the quality of summarization in terms of accuracy, clarity and preservation of key information in the specific context of financial texts.
- Identifying the strengths and weaknesses of the models for processing terminology and analysis-intensive technical texts.
- Proposing recommendations for the effective use of language models in practice, including the possibility of adapting them to domain-specific tasks such as summarizing technical articles in the financial sector.

The results of the work will contribute to a deeper understanding of the capabilities of modern language models in specialized domains and to the design of NLP applications that will improve the processing and use of specialized financial information.

Research questions

- Can local solutions for abstract summarization of financial text match the performance of global models?
- What are the fundamental differences between the models, both in clarity and in the ability to identify key information?
- Which of the models tested best fits the specific needs of summarizing financial articles?

2 LITERATURE REVIEW

2.1 Text summarization

Text summarization endeavors to produce a summary version of a text, while maintaining the original ideas (Nazari and Mahdavi, 2018). According to Liu *et al.* (2023), text summarization is an important NLP task where the goal is to generate a shorter version of an input text while preserving its main ideas. Generally, it is a process of creating a shorter version of a longer text that still contains the main ideas and key information. The goal of summarization is to allow the reader to quickly understand what the text is about without having to read everything. Instead of reading every word, summarization can pick out the most important parts – like main facts, key points, or important conclusions. The summary should have good structure, and the sentences should be coherent (Yadav *et al.*, 2022). The result is a shorter text that is still understandable and effectively captures the original content. In essence, it's about "extracting the essence" from a long text and packaging it into a shorter, simpler version. According to (Raman and Meenakshi, 2020), Text summarization has now become the need for numerous applications, like market review for analysts, search engine for phones or PCs, business analysis for businesses. One of the main approaches, when viewed from the summary results, are extractive and abstractive (Widyassari *et al.*, 2022).

- **Abstractive text summarization:** summarizing using the model‘s own formulations (Maylawati *et al.*, 2024, Yang *et al.*, 2020, Sinha *et al.*, 2018).
- **Extractive text summarization:** summarizing by directly selecting important sentences from the text (Paulos *et al.*, 2024, Rahman *et al.*, 2021).
- **Hybrid text summarization:** It combines both extractive and abstractive methods. It means extracting some sentences and generating a new one from a given corpus (Binwahlan *et al.*, 2010).

The evaluation of text summarization approaches has undergone a dynamic evolution, progressing from traditional metrics to semantics-focused metrics and incorporating human evaluations (Supriyono *et al.*, 2024). In order to effectively summarize, syntactic, semantic, and pragmatic concerns become crucial, highlighting the necessity of capturing not only grammar but also the context and underlying meaning (Supriyono *et al.*, 2024). According to (Singh & Deepak, 2021, Sinha *et al.*, 2018), semantic, syntactic, and pragmatic considerations form the core of effective text summarization.

2.2 Types of financial articles

Overview articles are intended for general public and experienced investors. These articles summarize current developments in financial markets in order to provide readers with a comprehensive and easy-to-understand overview of key events, trends, or statistics (e.g., interest rate changes, corporate earnings) without deep analysis or forecasts. Example: "Monthly Summary of Stock Market Developments."

Analytical articles are intended for advanced investors, economists and professionals. These articles delve deeper and analyse specific phenomena, trends, or companies based on data and conclusions to give readers insights into the causes and consequences of specific situations, offer justified forecasts, data analyses, graphs, models, interpretations of financial indicators, or make recommendations. Example: "The Impact of ECB Monetary Policy on European Bond Markets."

Technical articles are intended for academics, quantitative analysts, or technical finance experts. These articles focus on the technical aspects of finance, such as mathematical models, algorithms, data analysis methods, or technical details of financial instruments in order to explain, develop specific technical methods or approaches or describe of models, algorithms, case studies. Example: "Portfolio Optimization Using the Markowitz Model in Python."

3 METHODOLOGY AND DATA

It is difficult for people to recognize what information should be included in a summary; therefore, evaluating it is difficult (Yadav *et al.*, 2022). Also (Gambhir & Gupta, 2017) confirm that information changes depending on the summary's purpose, and mechanically capturing this information is a challenging undertaking.

The subject of our research was the ability of the selected AI models to summarize financial articles. Specifically, the versions GPT-4o, Mistral Instruct, and Llama 3.1 8B Instruct 128k were used. The capabilities of the GPT model were tested in the ChatGPT 4o interface, while the capabilities of the Mistral and Llama models were tested in the GPT4All environment, version 3.2.1.

3.1 Data preparation

In the first stage, a representative set of financial articles of varying length, complexity and style were selected. These articles were obtained from public web sites focused on financial markets like Bloomberg[1], SeekingAlpha[2] or YahooFinance[3]. We prepared a diverse dataset, including overview articles (O), analytical articles (A) and technical texts (T). We have also included articles of varying lengths to assess each model's ability to process different volumes of information, ranging approximately from 20 to 6,000 words.

We focused mainly on the ABSTRACT SUMMARIZATION that means summarizing using the model‘s own formulations. The models‘ task was to generate a concise and informative summary from the given article. The summary should include important details while maintaining coherence and clarity. For all three models, we set up the same output requirements:

Tab. 1 Output requirements

REQUIREMENT	EXPLANATION
Output format	JSON format with valid, iterable RFC8259 compliant code in your responses
Number of JSON fields	1. summary field with text containing summary of the article 2. bulletpoints field with list of strings containing three main giveaway from the article in a form of short bulletpoints
Length of the summary	Only 5 sentences
Length of the bulletpoints	Only 3 items

3.2 Evaluation criteria

The two most significant aspects of judging a summary are its quality and informativeness (Yadav, 2022). Our evaluation relies on qualitative methodologies combined with manual analysis to ensure a thorough and detailed assessment of the models. Specifically, we concentrated on the following key aspects:

- **Ability to handle text of different lengths** – this criterion evaluates the model's ability to generate coherent and relevant summaries for inputs of varying lengths, from short texts (20–100 words) to longer documents (over 5000 words).

[1] https://www.bloomberg.com/europe

[2] https://seekingalpha.com/

[3] https://finance.yahoo.com/

- **Ability to generate JSON code correctly** – this criterion assesses the model's accuracy in generating properly structured JSON code when required. The evaluation will check whether the output is syntactically valid JSON and whether it includes all required key-values.
- **Ability to meet output length requirements** – this criterion measures whether the model adheres to specific formatting requirements for the output—specifically, generating summaries of exactly 5 sentences and 3 bullet points. The evaluation will be conducted by counting the number of sentences and bullet points in the output.
- **Summarization efficiency** – specifically, we focused on:
  - **Total response length** – this metric captures the overall length of the model's entire response in words.
  - **Length of summary** – this metric measures the length of the generated summary in words.
- **Compression ratio of summary** is calculated as:

Compression Ratio (Summary)= Total Length of SummaryLength of Input

  - **Length of bullet points** – this metric evaluates the total word count of the bullet points generated by the model.
  - **Compression ratio of bullet points** is calculated as:

Compression Ratio (Bullet Points)= Total Length of Bullet PointsLength of Input

- **Quality of summarization**:
  - **Accuracy of information** – measures whether the summary captures the key points and is factually accurate, with a high degree of alignment with the essential information in the original article.
  - **Clarity of message** – assesses the quality of the language, readability, and clarity of the summary.
  - **Relevance of information** – evaluates the model‘s focus on essential information (e.g., the main ideas and arguments of the article).

3.2.1 Rating scale for the summarization quality criteria

For evaluating the quality of summarization, a five-point rating scale was employed For evaluating the quality of summarization, a five-point rating scale was employed across three key criteria: **accuracy**, **clarity**, and **relevance**. The scale allows to compare the performance of different AI models in producing high-quality summaries. Below is a detailed explanation of the meaning of each item on the scale:

1. (Very Poor) – The model fails to meet the basic requirements of the criterion.
2. (Poor) – The model meets some requirements but has significant shortcomings.
3. (Average) – The model meets the basic requirements but lacks precision or consistency.
4. (Good) – The model meets most of the criterion's requirements with minor issues.
5. (Excellent) – The model fully meets all requirements without significant errors.

This five-point scale ensures that both quantitative analysis (through numerical scores) and qualitative analysis (based on detailed human evaluation) are considered in the overall assessment. The final score for summarization quality is determined by averaging the scores across the three criteria (accuracy, readability, and relevance). The following table summarizes the detailed scoring methodology for each of the summarization quality criteria.

Tab. 2 Detailed scoring methodology for each of the summarization quality criteria

POINTS	INFORMATION ACCURACY	CLARITY OF THE MESSAGE	RELEVANCE OF INFORMATION
1	The summary contains inaccuracies, distortions, or misses key points.	The language is unclear, full of errors, and difficult to read.	The summary primarily includes irrelevant or unrelated information.
2	Some key points are captured, but many important details are missing or distorted.	The language has frequent errors, and the summary is unclear or poorly organized.	Includes a few relevant points but still has excessive irrelevant or redundant content.
3	Most key points are accurate, but some details may be incorrect or omitted.	The language is readable but occasionally unclear or stylistically inconsistent.	Focuses on main ideas but includes some unnecessary or secondary information.
4	The summary is factually correct and captures nearly all key points.	The language is clear, readable, and stylistically appropriate.	Focuses on essential information with minimal extraneous details.
5	The summary accurately and comprehensively captures all key points of the article.	The language is highly understandable, stylistically polished, and well-structured.	The summary focuses entirely on the main ideas and arguments, with no superfluous details.

3.3 Experimental setup

First, we tested a small number of articles to evaluate how the models behave and to adjust the settings. To obtain statistically valid results, we selected a sufficient number of articles depending on the variability of the texts. For the evaluation, we used qualitative scores to provide a comprehensive view.

3.4 Testing principles

In particular, we always started each conversation with an AI model in a new chat, as the model only retains context within the current chat. Starting a new chat ensures that the model is not influenced by previous context, minimizing the risk of misinterpreting the question or biasing the answer. It also improves the accuracy and consistency of the output, because the model is working with clean and unambiguous input. The conversation always started with the sentence setting the context for the model: "*You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.*" This formulation improves the consistency and correctness of the answer.

4 RESULTS AND DISCUSSION

A total of 30 articles from the field of finance were analyzed in detail, ensuring a diverse representation of financial topics, such as corporate finance, investment strategies, and market analysis. This sample size is sufficient to capture variability in content while maintaining a manageable scope for thorough evaluation. Additionally, this number allows for basic statistical analysis, enabling comparisons and trend identification across models. A larger sample could have reduced the depth of evaluation, while a smaller one might not have provided enough variability for reliable conclusions.

4.1 Ability to handle text of different lengths

During testing, the GPT model successfully processed all inputs, including those up to 6000 words. In contrast, the Mistral and Llama models frequently encountered input length issues, returning errors indicating that the prompt size exceeded their respective context windows. Specifically, the Mistral model refused to process inputs longer than approximately 1200 words, while the Llama model could not handle inputs exceeding 1400 words. The context window sizes reported by the models varied significantly: Mistral indicated a limit of 10,240 characters, Llama reported a maximum prompt size of 2048 characters (approximately 300–400 words), while GPT claimed a substantially larger capacity of approximately 100,000 tokens, encompassing both input and output within a conversation.

Notably, the texts that Mistral and Llama failed to process were predominantly analytical in nature. These texts involved complex arguments, detailed explanations, and extensive use of data, which likely contributed to their increased length. This highlights a potential limitation of Mistral and Llama when handling information-dense, structured inputs, particularly in contexts requiring detailed analytical thinking.

For consistency in evaluation, only texts for which all three models produced responses were included. This approach ensured a balanced comparison by focusing exclusively on cases where the output from all models could be analyzed.

4.2 Ability to generate JSON code correctly

The evaluation of the models' ability to generate JSON code correctly revealed that both the Llama and GPT models consistently produced valid and error-free JSON outputs. In two instances, the Mistral model failed to include the trailing parenthesis, resulting in syntactically invalid JSON code, which could not be parsed correctly.

GPT consistently generates well-formatted JSON code with proper indentation, ensuring a clear and organized structure. This improves readability and facilitates better comprehension of the logical flow, particularly in complex outputs. In contrast, both Mistral and Llama produce JSON outputs without proper indentation, presenting the code as a single continuous block of text. This lack of formatting complicates readability, making it more difficult for users to discern key elements or understand the hierarchical structure of the data.

Proper formatting, such as indentation, is essential for improving the user experience, particularly in scenarios involving complex JSON structures where readability and ease of debugging are critical.

4.3 Ability to meet output length requirements

All three models consistently met the requirement of generating three bullet points for each text tested. However, there were notable differences in meeting the requirement to produce exactly 5 sentences of summarization. The GPT model fully met the specification, consistently generating 5 sentences in each case. In contrast, the Llama model deviated on three occasions, producing 4 sentences instead of 5, which may reflect minor challenges in sentence segmentation or adherence to the specification.

The Mistral model showed the greatest inconsistency, failing to meet the 5-sentence requirement in 18 of the cases tested. The number of sentences generated by Mistral varied significantly, ranging from 1 to 7 sentences. This variability suggests potential limitations in Mistral‘s ability to consistently follow output length constraints, and highlights a need for improvement in handling precise output formatting requirements.

An additional test was conducted using an edge case, where the financial article consisted of a single informative sentence containing 19 words. The GPT model adhered strictly to the specified requirements, generating 5 sentences and 3 bullet points. However, it lengthened the original text by expanding the single sentence into multiple sentences rather than condensing it. The Llama model generated the required 3 bullet points but produced only 3 sentences instead of the specified 5, indicating a tendency to condense content.

We also tested the extreme case where the financial article was a single informative sentence of 19 words. The GPT model stuck strictly to its specification and again generated 5 sentences and 3 bullets, lengthening the text instead of shortening it. The Llama model also generated 3 bullets, but shortened the summary to 3 sentences. The Mistral model was the only one to generate a one-sentence summary, dropping 3 irrelevant words from the original sentence. However, it still generated 3 bullet points. The Mistral model was the only one to produce a one-sentence summary, effectively shortening the original input by omitting 3 irrelevant words. Despite its brief summary, it still fulfilled the requirement by generating 3 bullet points, demonstrating its capacity to condense content while adhering to the specified output format.

4.4 Summarization efficiency

In this section, we evaluated summarization efficiency by measuring the average length of the total responses, the summaries, and the bullet points in terms of word count. The average length of the total responses varied across models. The Llama model produced the shortest output with an average of 118 words, followed by the Mistral model with 138 words. The GPT model generated the longest output with and average of 141 words. These results suggest that Llama prioritized brevity, while GPT tended to provide more detailed responses.

Regarding the length of the generated summaries, Mistral produced the shortest summaries, averaging 72 words, followed by Llama with 83 words. GPT generated the longest summaries, averaging 98 words. This indicates that GPT provided more detailed summaries, while Mistral prioritized conciseness.

The Llama model generated the shortest bullet points, averaging 32 words, followed by GPT with 36 words. Mistral produced the longest bullet points, averaging 61 words. Shorter bullet points, such as those generated by Llama, are generally more readable and suitable for quick information retrieval, while longer bullet points, like those from Mistral, may reduce clarity.

The compression ratio, which measures the degree of output compression relative to the input length, was also evaluated. For summarization, Mistral achieved the highest compression (0.187), followed by Llama (0.278), while GPT had the lowest compression (0.341), corresponding to its longer average output. A similar trend was observed for bullet point compression, with Llama achieving the highest compression ratio (0.101), followed by GPT (0.138), and Mistral producing the least compressed bullet points (0.195)

4.5 Quality of summarization

The summarization quality of all three AI models was evaluated based on three key criteria: **accuracy of information, clarity of message**, and **relevance of information**. Each criterion was rated on a five-point scale, where 1 represents the lowest quality and 5 represents the highest.

Overall, the summarization quality of the AI models was found to be very high, with an average score of 4.36 across all three criteria and models. This is also confirmed, for example, by a study of Goriparthi (2021) which states significant improvements in both summarization and translation quality of AI models. Our results indicate that, despite minor differences, the models are generally capable of producing high-quality summaries. Among the three criteria, the highest average score was achieved in **clarity of message** (4.85), followed by **relevance of information** (4.19) and **accuracy of information** (4.13). This suggests that while all models excel at presenting information clearly, slight inaccuracies may still occur.

Tab. 3 Overall results of the quality of summarization

	MISTRAL	LLAMA	GPT
Accuracy of information	3.81	4.04	4.54
Clarity of message	4.65	4.92	4.96
Relevance of information	7.73	4.12	7.72
Total score	4.06	4.36	4.74

The average scores for each model across the three criteria are as follows:

- Accuracy of information: Mistral (3.81), Llama (4.04), GPT (4.54)
- Clarity of message: Mistral (4.65), Llama (4.92), GPT (4.96)
- Relevance of information: Mistral (3.73), Llama (4.12), GPT (4.73)

These results show that GPT consistently outperformed the other models, particularly in accuracy and relevance, while all three models excelled in clarity. While there are some differences in performance, all three models demonstrated a high capacity for summarization, with even the lowest-performing model achieving an overall score above 4.0. This indicates that current AI models are well-suited for generating clear and relevant summaries in the field of finance. The following table brings the overall results of the quality of summarization.

4.6 Overall evaluation

The following table presents a comprehensive summary of the evaluation results for all three AI models. In addition to the core criteria outlined in the methodology, two additional factors were identified during testing as critical for practical application: **response speed** and **visual formatting of code**.

Tab. 4 Overall evaluation

EVALUATION CRITERION	MISTRAL	LLAMA	GPT
Length of input	*	**	***
JSON code correctness	**	***	***
Requirements for output	*	**	***
Total response length	**	***	*
Length of summary	***	**	*
Compression ratio of summary	***	**	*
Length of bullet points	*	***	**
Compression ratio of bullet points	*	***	**
Accuracy of information	*	**	***
Clarity of message	**	***	***
Relevance of information	*	**	***
Response speed	**	**	***
Visual formatting of code	*	*	***

While the main focus of the evaluation was on summarization quality, response speed emerged as a significant criterion, with GPT consistently outperforming both the Mistral and Llama models. Fast response times are essential in real-world applications, particularly in environments where quick information retrieval and decision-making are required. A slow response may hinder user productivity, especially in scenarios involving large datasets or time-sensitive analyses.

Another important factor beyond the core methodology was the visual formatting of generated code. GPT demonstrated superior formatting, producing well-indented and structured code, whereas Mistral and Llama often generated code as a single continuous block of text. Proper formatting enhances readability and facilitates easier debugging and integration of the generated code into existing systems. This is especially critical for developers and analysts working with structured data, as poorly formatted code increases the cognitive load and the likelihood of errors.

These additional criteria highlight that beyond pure summarization quality, practical usability factors such as speed and output clarity play a vital role in determining the overall performance of AI models in real-world tasks.

The results showed that GPT consistently outperformed both Mistral and Llama in terms of overall summarization quality, achieving the highest scores in accuracy, clarity, and relevance. Specifically, GPT excelled at producing clear, concise, and well-structured summaries with minimal errors in information accuracy. It also demonstrated superior adherence to output formatting requirements, consistently producing summaries of the specified length and properly structured JSON code with excellent visual formatting.

While Llama demonstrated competitive performance, particularly in terms of message clarity and summarization efficiency, it occasionally struggled with adherence to strict length requirements, sometimes generating fewer sentences than specified. Mistral, while the least consistent of the three models, demonstrated strengths in conciseness and compression ratio, but showed significant variability in meeting output requirements and generating error-free JSON code.

An additional aspect evaluated during the study was the practical usability of the models, specifically response speed and visual formatting of the generated code. GPT once again outperformed its competitors in these areas, providing the fastest response times and the most user-friendly output formatting. These factors are critical in real-world applications where clarity of output and fast response times are essential for efficient decision making and information retrieval.

5 CONCLUSION

This research highlights the significant potential of advanced AI models for financial summarization tasks, with GPT emerging as the most capable model in terms of quality, usability, and adherence to specified output requirements. However, the study also underscores the need for further improvements in model consistency, especially for Mistral and Llama, and suggests several directions for future research that could enhance AI summarization capabilities in specialized domains. As AI models continue to evolve, their application in financial analysis and other high-stakes fields is likely to become increasingly prominent, making rigorous evaluation frameworks such as the one used in this study essential for guiding their development and deployment.

REFERENCES

BINWAHLAN, M. S., SALIM, N., SUANMALI, L. 2010. Fuzzy swarm diversity hybrid model for text summarization. *Information Processing and Management*, 46(5), 571–588. https://doi.org/10.1016/j.ipm.2010.03.004

DHAINI, M., ERDOGAN, E., BAKSHI, S., KASNECI, G. 2024._Explainability Meets Text Summarization: A Survey. In: *Proceedings of the 17th International Natural Language Generation Conference*. Tokyo, Japan: Association for Computational Linguistics, p. 631–645.

GAMBHIR, M., GUPTA, V. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*. 47(1), 1–66. https://doi.org/10.1007/s10462-016-9475-9

GORIPARTHI, R. G. 2021. AI-Driven Natural Language Processing for Multilingual Text Summarization and Translation. *Revista de Inteligencia Artificial en Medicina*. 12(1), 513–535. [Accessed 2025, January 11] http://redcrevistas.com/index.php/Revista/article/view/226

LIU, Y. L., CAO, M., BLODGETT, S. L., CHEUNG, J. C. K., OLTEANU, A., TRISCHLER, A. 2023. Responsible AI Considerations in Text Summarization Research: A Review of Current Practices. arXiv preprint. *arXiv*: 2311.11103v1. https://doi.org/10.48550/arXiv.2311.11103

MAYLAWATI, D. S., SHALIH, K. M., RAMDHANI, M. A., SLAMET, C., RAMDANIA, D. R. 2024. Indonesian Abstractive Text Summarization with Bidirectional Long Short-Term Memory (Bi-LSTM). In: 12th International Conference on Cyber and IT Service Management (CITSM). https://doi.org/10.1109/CITSM64103.2024.10775593

NAZARI, N., MAHDAVI, M. A. 2018. A survey on Automatic Text Summarization. *Journal of AI and Data Mining*. 7(1), 121-135. https://doi.org/10.22044/jadm.2018.6139.1726

PAULOR, E. B., WOLDEYOHANNIS, M. M., DANA B. S., YESUF S. M., YIGEZU M. G. 2024. Extractive Text Summarization for Wolaytta Language Using Recurrent Neural Network. In International Conference on Information and Communication Technology for Development for Africa (ICT4DA). https://doi.org/0.1109/ICT4DA62874.2024.10777225

RAMAN, J., MEENAKSHI, K. 2021. Automatic Text Summarization of Article (NEWS) Using Lexical Chains and WordNet—A Review. In: HEMANTH, D., VADIVU, G., SANGEETHA, M., BALAS, V. (Eds.). *Artificial Intelligence Techniques for Advanced Computing Applications*. Lecture Notes in Networks and Systems, vol 130. Singapore: Springer. https://doi.org/10.1007/978-981-15-5329-5_26

RAHMAN, N. A., ; RAMLAM, S. N. A., AZHAR, N. A., HANUM, H. M., RAML, N. I. AND LATEH, N. 2021. Automatic Text Summarization for Malay News Documents Using Latent Dirichlet Allocation and Sentence Selection Algorithm. In: *Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*. IEEE. https://doi.org/10.1109/CAMP51653.2021.9498029

ROY, K., MUKHERJEE, S., DAWN, S. 2023. Automated Article Summarization using Artificial Intelligence Using React JS and Generative AI. *Journal of Emerging Technologies and Innovative Research (JETIR)*. 10(6), 78–87. ISSN: 2349-5162.

SAIYYAD, M. M., PATIL, N. N. 2024. Text Summarization Using Deep Learning Techniques: A Review. *Engineering Proceedings*. 59(1), 194. https://doi.org/10.3390/engproc2023059194

SINGH, S., DEEPAK, G. 2021. Towards a Knowledge Centric Semantic Approach for Text Summarization. In: *Data Science and Security*. Lecture Notes in Networks and Systems, vol 290. Singapore: Springer. https://doi.org/10.1007/978-981-16-4486-3_1

SINHA, A., YADAV, A., GAHLOT, A. 2018. Extractive Text Summarization using Neural Networks (Version 1). *arXiv*. 1802.10137. https://doi.org/10.48550/ARXIV.1802.10137

SUPRIYONO, WIBAWA, A. P., SUYONO, KURNIAWAN, F. 2024. A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*. 7, 100070. Elsevier BV. https://doi.org/10.1016/j.nlp.2024.100070

WIDYASSARI, A. P., RUSTAD, S., SHIDIK, G. F., NOERSASONGKO, E., SYUKUR, A., AFFANDY, A., SETIADI, D. R. I. M. 2022. Review of automatic text summarization techniques and methods. *Journal of King Saud University – Computer and Information Sciences*. 34(4), 1029–1046. Springer Science and Business Media LLC. https://doi.org/10.1016/j.jksuci.2020.05.006

YADAV, D., DESAI, J., YADAV, A. K. 2022. Automatic Text Summarization Methods: A Comprehensive Review (Version 1). *arXiv:*2204.01849v1. https://doi.org/10.48550/ARXIV.2204.01849

YANG, M., WANG, X., LU, Y., LV, J., SHEN, Y., LI, C. 2020. Plausibility-promoting generative adversarial network for abstractive text summarization with multi-task constraint. *Information Sciences*. 521, 46–61. https://doi.org/10.1016/j.ins.2020.02.040

Acknowledgement

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-founded by the Operational Programme Research, Development and Education, and by the project OP TAK 2023-2025 Analytical Platform 2.0 – automated portfolio management platform using learning systems.

Contact information

Jana Dannhoferová: e-mail: jana.dannhoferova@mendelu.cz
Jan Přichystal: e-mail: jan.prichystal@mendelu.cz