ISSN 2224-087X. Електроніка та інформаційні технології. 2024. Випуск 28. С. 3–13
Electronics and information technologies. 2024. Issue 28. P. 3–13

UDC 004.932 DOI: https://doi.org/10.30970/eli.28.1

COMPARISON OF ZERO-SHOT APPROACH AND RETRIEVAL-AUGMENTED GENERATION FOR ANALYZING THE TONE OF COMMENTS IN THE UKRAINIAN LANGUAGE

M. Prytula, O. Sinkevych, I. Olenych

Radielectronic and Computer Systems Department,
Ivan Franko National University of Lviv,
50 Drahomanova St., UA–79005 Lviv, Ukraine
marianna.prytula@lnu.edu.ua , *oleh.sinkevych@lnu.edu.ua* , *igor.olenych@lnu.edu.ua*

Background. The constant growth of information, online news and text messages in social networks causes new challenges for society. It requires robust tools for analyzing information in real-time, including determining its emotional tone. Understanding the emotional aspect directly affects customer satisfaction in various areas of activity and can suggest directions for improving processes. Therefore, the development of tools for analyzing the tonality of texts can provide the ability to accurately recognize people's emotions, identify problems, and determine ways to solve them.

Methods. In this study, approaches to the application of the Mistral-7B-UK large language model were implemented for the text tone analysis. Two datasets of comments in the Ukrainian language were utilized: one for binary classification, divided into negative and positive classes, and another for multiclass classification which included a neutral tonality. These datasets contain reviews about shops, restaurants, hotels, medical facilities, entertainment centers, fitness clubs, the provision of various services, etc.

Results and Discussion. The prompts were constructed for the zero-shot approach, describing the role, output format, and additional explanation about tonalities. To implement RAG, Qdrant was utilized as a vector database, while the LangChain library enabled the integration of a large language model with external data sources. To determine text tonality, the five most semantically similar chunks with the defined tonality are retrieved from the vector database, and predefined placeholders are filled in the prompt template. The model's response is generated using the provided context.

Conclusion. Research showed that the zero-shot approach achieves higher text tone analysis accuracy than the Retrieval-Augmented Generation model. For binary classification, the overall accuracy was 94 %, and for multiclass – 75 %. The benefit of using external sources was found during the model's recognition of neutral tonality. However, it was observed that comments with opposing tonality could be retrieved as context due to the shared object of description, which negatively affects results.

Keywords: text tone, Large Language Model, zero-shot, Retrieval-Augmented generation.

Introduction

The continuous growth of Internet-based applications, such as e-commerce websites, online forums, and social media platforms, has led to the generation of a large volume of textual information. Analyzing the tone of a text can be highly beneficial in several cases, as it helps to determine its emotional component, understand the audience's needs, and respond effectively to their expectations. This process is valuable for businesses, researchers, and marketers. For example, qualitative and prompt responses to detected negative feedback provide a higher customer service level, evaluation of people's reactions outlines the business improvements' paths, and analyzing current people's insights helps to fill an understanding gap between the government and the public so they can be on the same page about essential topics [1–4]. Therefore, the ongoing monitoring of text tonality allows for enhancement processes in different operation fields, emphasizing the necessity for developing robust classification tools.

In the era of Generative AI, large language models are widely used in natural language processing tasks, such as automatic translation, chatbots, tonality analysis, text classification, and content creation due to the ability to process and generate high-quality texts. LLMs can understand context, tone, and field-specific knowledge based on the extensive amounts of data they are trained on. These capabilities are essential for enhancing efficiency in numerous business and research tasks.

Several approaches can be used for text classification using large language models. Fine-tuning represents one of the strategies that has proven effective across various NLP tasks, allowing pre-trained knowledge to be adapted and fine-tuned for the target domain after additional training on specialized datasets. In research work [5], the BERT Base Uncased model fine-tuning has been presented to solve sentiment analysis tasks compared with Naive Bayes Classification, LSTM, and Support Vector Machine methods. The results have shown that BERT's effectiveness is much higher than other machine learning algorithms. In the paper [6], fine-tuning of BERT, DistilBERT, XLM-RoBERTa, and Ukr-RoBERTa models for sentiment analysis of reviews in the Ukrainian language was performed, and according to evaluation metrics, XLM-RoBERTa model has achieved the highest accuracy score.

Applying the zero-shot method means a large language model performs a new task without prior training on specific data. In contrast, few-shot training involves the transfer of several examples before solving the task, increasing the accuracy of answers based on the minimum amount of data. The article [7] represents a complex study for sentiment analysis on 20 online review datasets focusing on binary and three-class classification tasks. The study benchmarks GPT-3.5, GPT-4, and Llama 2 against fine-tuned variants of BERT-based transfer learning models. GPT-4 performed exceptionally well in zero-shot sentiment analysis, achieving 93% accuracy in binary classification and 83% in three-class classification. These results are comparable to or even better than some fine-tuned models, except for SiEBERT in the binary task. In the paper [8], the effectiveness of the domain adaptation approach to pre-training and fine-tuning BERT models is analyzed compared to zero-shot and few-shot using the GPT-3 model. The results have indicated that RoBERTa fine-tuned for specific classification tasks significantly outperforms both prompt-based learning methods using GPT-3. Another study [9] also compares LLMs in combination with zero-shot and one-shot methods to smaller but fine-tuned language models for text classification. The need for training data has been highlighted, as fine-tuning smaller language models has outperformed the in-context learning methods of larger text generation models.

Utilizing Retrieval-Augmented Generation is another way to employ large language models. The approach integrates retrieval mechanisms with generative models to improve the accuracy and relevance of information in natural language generation tasks. The process begins by retrieving pertinent documents or information fragments from an external knowledge base, which then influences the response produced by a language model. RAG is widely used for question-answering tasks, allowing the model to access up-to-date and specific information beyond its pre-trained knowledge. RAG models can create more accurate outputs by incorporating additional context within the generation process. This capability makes them extremely helpful when implementing tools for customer support, information retrieval systems, and interactive AI systems. Several studies [10, 11] have delved into investigating the Retrieval-Augmented generation. The article [12] highlights the advantages of allowing models to access external data sources during answer generation while evaluating QA systems. A Retrieval-Augmented LLM framework for financial sentiment analysis was introduced in the study [13]. Instruction tuning was implemented to enhance the performance of the Llama-7B model, allowing the LLM to predict sentiment labels effectively. Additionally, a retrieval-augmentation module was developed to fetch extra context from external sources. This approach demonstrated a performance improvement of 15% to 48% in accuracy and F1-score compared to traditional models and other LLMs, such as ChatGPT and LLaMA. The article [14] states that although retrieval-augmented generation methods are well-studied for knowledge-intensive tasks, their potential for non-knowledge-intensive tasks, such as sentiment analysis, text classification, and linguistic acceptability, still needs to be explored. It was found that sentences with thematically similar words may be retrieved, but they can convey opposing tones or fail to align with the intended task. This inconsistency can hinder an accurate classification of the text's tone. Consequently, the effect of RAG on improving results in non-knowledge-intensive tasks, particularly sentiment analysis, still requires further research and experimentation.

In the article [15], Chain of Thought prompting was explored in a range of arithmetic, common sense, and symbolic reasoning tasks, and it has shown performance improvement and expansion of the capabilities of large language models. This approach involves model training using detailed, step-by-step explanations in prompts to stimulate critical thinking and assist the model in generating structured and logical answers for complex tasks.

While text tone analysis is commonly categorized into positive and negative labels, the challenges associated with multiclass classification, especially in the Ukrainian language, inspire an interest in further study. Due to the contradictory results in articles that compare different methods for utilizing LLMs in specific domains, this research aims to identify a more effective approach for text tone classification. Therefore, the exploration and evaluation of zero-shot learning and Retrieval-Augmented generation were conducted for binary and multiclass analysis of multi-domain comments in the Ukrainian language.

Methods.

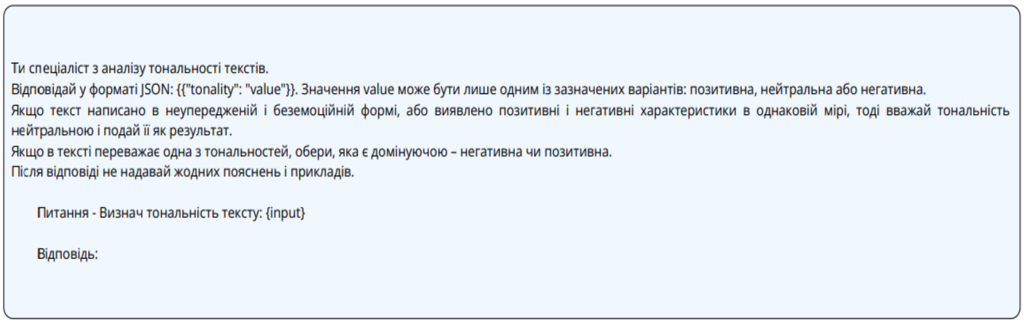
The research uses two datasets of comments in the Ukrainian language about shops, restaurants, hotels, medical facilities, entertainment centers, fitness clubs, the provision of various services, etc. Three labels are allocated for multiclass classification: negative, neutral, and positive. The textual data contained an even distribution of instances by class – 670 in each. Negative and positive tonalities were used for binary classification. The dataset includes 2,400 comments with an equal number of examples per class.

In the paper [16] different language models were tested for Ukrainian tasks, including GPT-3.5-turbo, GPT-4-1106-preview, several versions of Mistral, and SherlockAssistant/Mistral-7B-InstructUkrainian showed the highest results among non-OpenAI models for all tasks, even outperforming GPT-3 for some of them. This study shows that the use of Mistral-7B is promising for performing tasks in the Ukrainian language context. Therefore, the Mistral-7B-InstructUkrainian [17] was utilized to analyze text tone classification as a large language model.

Initially, zero-shot prompting was examined to assess the effectiveness of using LLM to analyze comments. According to this approach, no additional training or task-specific tuning is involved, as the primary assumption is that the model can accurately understand the instructions or context of the task-based solely on the request provided.

The prompt has contained a definition of the role, namely that of a specialist in analyzing the tonality of texts. The response format the model should respond to was defined as JSON, and some additional explanations about labels were added. After revising the comments labels and trying to identify regularities in the text, it was determined that neutral-colored comments are expressed in an unbiased and unemotional manner or contain both positive and negative characteristics in equal measures.

For multi-class classification, the prompt specifies the conditions under which texts with a neutral tone should be classified, as illustrated in Fig. 1. Also, guidance was provided that the dominant – negative or positive class should be determined if one of the tones prevails.


Ти спеціаліст з аналізу тональності текстів.
Відповідай у форматі JSON: {{"tonality": "value"}}. Значення value може бути лише одним із зазначених варіантів: позитивна, нейтральна або негативна.
Якщо текст написано в неупередженій і беземоційній формі, або виявлено позитивні і негативні характеристики в однаковій мірі, тоді вважай тональність нейтральною і подай її як результат.
Якщо в тексті переважає одна з тональностей, обери, яка є домінуючою – негативна чи позитивна.
Після відповіді не надавай жодних пояснень і прикладів.

Питання - Визнач тональність тексту: {input}

Відповідь:

Fig. 1. Prompt template used for multi-class classification.
Рис. 1. Шаблон підказки для багатокласової класифікації.

The next step was to integrate RAG to evaluate the effectiveness of text tone classification. The LLM model transforms data into high-dimensional vectors; for instance, embeddings of the used Mistral-7B have 4096 dimensions. Vector databases are used to provide efficient information storage and retrieval. They allow for the quick finding of semantically similar documents by comparing vectors in a multidimensional space. Several options for a vector database are applicable [18–19], such as Weaviate, Faiss, Chroma, Qdrant, Milvus, PostgreSQL, etc. A Qdrant database was chosen to implement the RAG model. The LangChain library was used to integrate large language models with the database.

For multi-class classification, 1,608 comments with an equal number for each label were randomly selected from the dataset. For binary classification, 1,000 documents each for the negative and positive tones were chosen. The documents were divided into chunks of a fixed length using a recursive character text splitter, each containing no more than 512 characters, with an overlap of 50 characters. These fragments were then converted into vectors using the Mistral-7B-InstructUkrainian model for storage in the Qdrant vector database. Additionally, a tonality label was added to each chunk as metadata. The cosine distance metric was used to perform similarity searches on vectors.

Fig. 2 illustrates the main steps to enhance language model responses by incorporating relevant external information. The pipeline starts with the user query, then converts it into a vector format and performs a semantic search in a vector database, which enriches the prompt with pertinent documents for the Large language model. Finally, the response is formed based on the general knowledge of the model and the provided context.

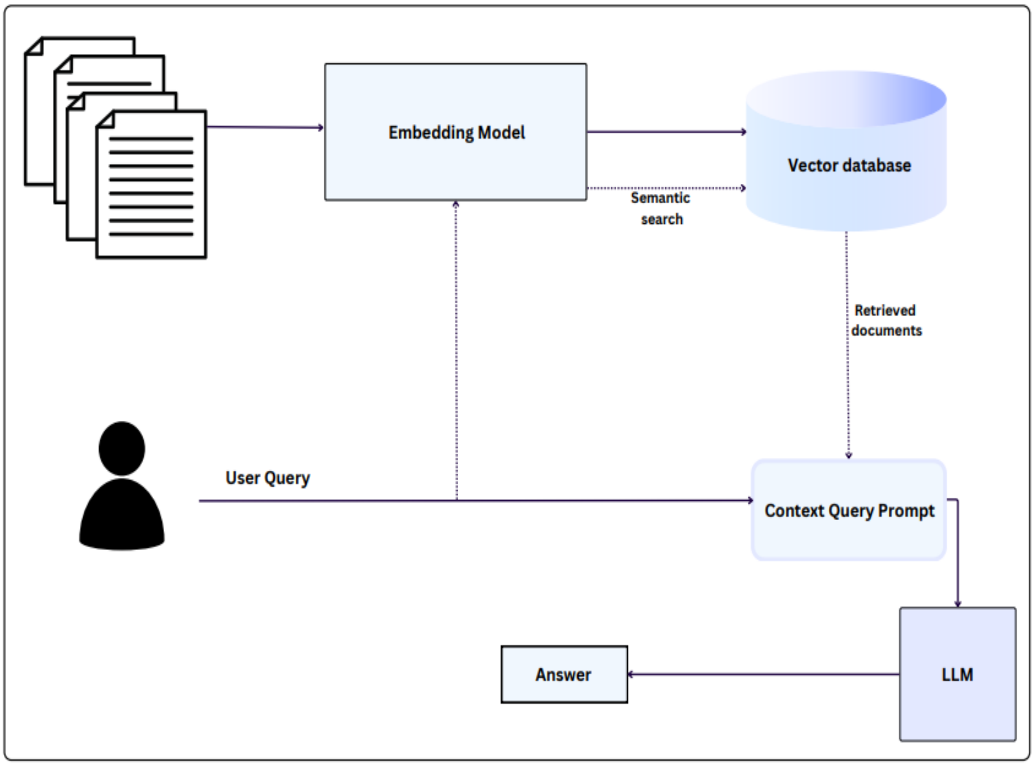


Fig. 2. Retrieval-Augmented Generation Model architecture.
Рис. 2. Архітектура моделі пошуку з доповненою генерацією.

Fig. 3 shows the prompt for binary classification, which starts by defining the role of a specialist in analyzing the tonality of texts. Then, the instruction was to use context and tonality values of chunks to answer the question. The response format the model should respond to was also defined as JSON. Additional guidance was specified that if the text contains both positive and negative tones, the dominant one should be identified.

```
Ти спеціаліст з аналізу тональності текстів.
Використовуй контекст та значення тональностей фрагментів, щоб відповісти на запитання.
Відповідай у форматі JSON: {{"tonality": "value"}}, де значення value може бути тільки - позитивна або негативна.
Якщо в тексті виявлено одночасно і позитивну, і негативну тональності, тоді визнач, яка з них є домінуючою і подай її як основний результат.

  Питання - Визнач тональність тексту: {input}
  Контекст: {context}

  Тональність фрагментів: {tonality}

  Відповідь:
```

Fig. 3. Prompt template used for binary classification.using RAG.
Рис. 3. Шаблон підказки для бінарної класифікації з використанням пошуку з доповненою генерацією.

After entering, a query is converted into a vector format, and a similarity search is performed. The five most relevant chunks from the vector database are fetched and located in a context placeholder in the prompt template. The class labels from metadata are put in a tonality placeholder, and the user query is also added as input. The constructed prompt is transmitted as text to a large language model, generating a response based on relevant information from the database.

Results and discussion.

The zero-shot approach was tested on the remaining 402 comments for multi-class classification. According to the confusion matrix in Fig. 4, the Mistral-7B-UK model struggles to recognize neutral tonality. It correctly classified only 66 out of 134 neutral comments, and 32 samples from other totalities were mistakenly included in this category. However, the model performs almost excellently in distinguishing between positive and negative tones, as only one comment from each category was misclassified as belonging to the opposite tone.

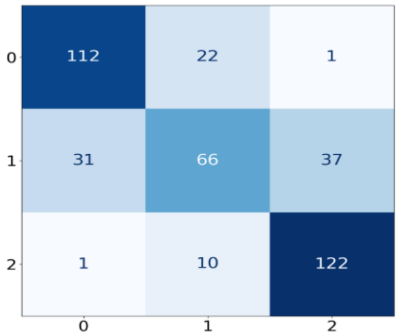


Fig. 4. Confusion matrix for multiclass classification using zero-shot approach.
Рис. 4. Матриця невідповідностей для багатокласової класифікації з використанням zero-shot підходу.

For binary classification testing, the sampling of 400 comments was used. Fig. 5. shows that only 8 out of positive comments were labeled as negative, and 16 positive comments were recognized as negative class.

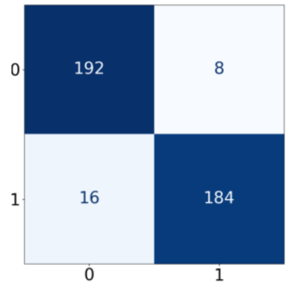


Fig. 5. Confusion matrix for binary classification using zero-shot approach.
Рис. 5. Матриця невідповідностей для бінарної класифікації з використанням zero-shot підходу.

The implemented RAG model was tested on the same sampling comments for multiclass classification. According to the confusion matrix (Fig.6), the analysis of adjacent tonalities also causes difficulties for the model since a significant number of neutral comments are entered into the negative and positive classes and positive and negative ones into the neutral class. However, the number of misclassified true neutral comments is lower than the zero-shot approach. This suggests that external sources have positively influenced the understanding of how people label the neutral category. Conversely, a higher number of positive comments were categorized as negative and neutral. Upon analyzing the RAG model's answers, one reason for the false classification was revealed: comments with an opposite tone were selected as context due to describing the same object. For instance, if the input text expresses an opinion about a store or restaurant, the most similar retrieved chunks from the database may also contain information about these establishments, but convey contrasting emotions.

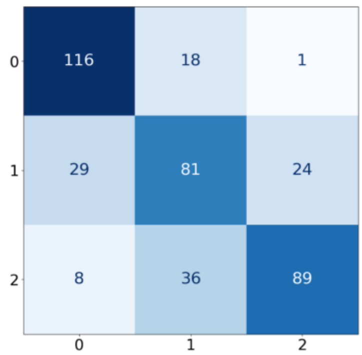


Fig. 6. Confusion matrix for multiclass classification using RAG.
Рис. 6. Матриця невідповідностей для багатокласової класифікації з використанням RAG.

The negative comments were easily recognized during binary classification using the RAG model, as only five were labeled as positive (Fig. 6). However, 56 positive documents were interpreted as negative.

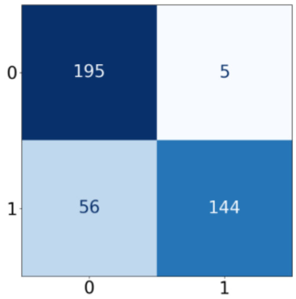


Fig. 7. Confusion matrix for binary classification using RAG.
Рис. 7. Матриця невідповідностей для бінарної класифікації з використанням RAG.

The results of the zero-shot approach and the RAG model's effectiveness in determining the text tone of comments divided into three classes are presented in the Table. 1.

Table 1. Evaluation metrics of utilizing the zero-shot approach and the RAG model for analyzing the tone of comments divided into three classes

	Class	**Precision**	**Recall**	**F1-score**
Zero-shot	Negative	78 %	83 %	80 %
	Neutral	67 %	49 %	57 %
	Positive	76 %	92 %	83 %
RAG	Negative	76 %	86 %	81 %
	Neutral	60 %	60 %	60 %
	Positive	78 %	67 %	72 %

Comparing the F1-score metric, which represents a harmonic mean of precision and recall, indicates that the RAG model has outperformed the zero-shot approach for the neutral class by 3% and negative tone by 1%; however, for the positive class, the zero-shot approach has achieved a result that was 11% higher.

Table 2. Evaluation metrics of utilizing the zero-shot approach and the RAG model in the binary classification of comments tonality

	Class	**Precision**	**Recall**	**F1-score**
Zero-shot	Negative	92 %	96 %	94 %
	Positive	96 %	92 %	94 %
RAG	Negative	78 %	97 %	86 %
	Positive	97 %	72 %	83 %

According to the results in Table 2 for binary classification, the zero-shot approach also demonstrated a higher F1-score equal to 94% for both classes.

Conclusion.

The study focused on applying the Mistral-7B-UK large language model for text tone analysis in the Ukrainian language. Considerate emphasis was placed on multiclass classification, which included negative, neutral, and positive categories. Additionally, an analysis of binary classification was also conducted.

A comparison was performed using the zero-shot approach versus the Retrieval-Augmented Generation to assess the text tone classification efficiency. The results indicated that the zero-shot approach has demonstrated a higher performance in analyzing the tone of comments. The overall accuracy was 94 % for binary classification and 75 % for multiclass. According to the datasets, accuracy exceeded that of the RAG model by 9 % and 4%, respectively. The advantage of using external sources was particularly evident during multiclass classification for neutral tonality. However, during the analysis of RAG model answers, it was found that comments of the opposite tonality could be retrieved as a context due to a shared object of description.

Compliance with ethical standards.

The authors declare that they have no competing interests.

Author contributions.

Conceptualization, [MP, IO]; methodology, [MP, OS]; investigation, [MP, OS]; writing – original draft preparation, [MP]; writing – review and editing, [MP, IO].

All authors have read and agreed to the published version of the manuscript.

References

[1] *Nwosu, N. T., Babatunde, S. O., & Ijomah*, T. (2024). Enhancing customer experience and market penetration through advanced data analytics in the health industry. *World Journal of Advanced Research and Reviews*, 22(3), 1157–1170. https://doi.org/10.30574/wjarr.2024.22.3.1810

[2] *Idris, S. L., & Mohamad, M.* (2023). A Study on Sentiment Analysis on Airline Quality Services: A Conceptual Paper. *Information Management and Business Review*, 15(4), 564–576. https://doi.org/10.22610/imbr.v15i4(SI)I.3638

[3] *Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P. M.* (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223, 119862. https://doi.org/10.1016/j.eswa.2023.119862

[4] *Adeniran, I. A., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Agu, E. E., & Efunniyi, C. P.* (2024). Data-Driven approaches to improve customer experience in banking: Techniques and outcomes. *International Journal of Management & Entrepreneurship Research*, 6(8), 2797–2818. https://doi.org/10.51594/ijmer.v6i8.1467

[5] *Geetha, M. P., & Karthika Renuka, D*. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *International Journal of Intelligent Networks*, 2, 64–69. https://doi.org/10.1016/j.ijin.2021.06.005

[6] *Prytula, M.* (2024). Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews. *Artificial Intelligence*, 3(4), 85–97. https://doi.org/10.15407/jai2024.02.085

[7] *Krugmann, J.O., & Hartmann, J.* (2024). Sentiment Analysis in the Age of Generative AI. *Customer Needs and Solutions*, 11(1), 3. https://doi.org/10.1007/s40547-024-00143-4

[8] *Bosley, M., Jacobs-Harukawa, M., Licht, H., & Hoyle, A.* (2023). Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research. https://mbosley.github.io/papers/bosley_harukawa_licht_hoyle_mpsa2023.pdf

[9] *Edwards, A., & Camacho-Collados, J.* (2024) Language Models for Text Classification: Is In-Context Learning Enough? https://doi.org/10.48550/arXiv.2403.17661

[10] *Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A. & Grave, E.* (2023). Augmented language models: a survey. https://doi.org/10.48550/arXiv.2302.07842

[11] *Şakar, T., & Emekci, H.* (2024). Maximizing RAG efficiency: A comparative analysis of RAG methods. *Natural Language Processing*, 1–25. https://doi.org/10.1017/nlp.2024.53

[12] *Mansurova, A., Mansurova, A., & Nugumanova, A.* (2024). QA-RAG: Exploring LLM Reliance on External Knowledge. *Big Data and Cognitive Computing*, 8(9), 115–131. https://doi.org/10.3390/bdcc8090115

[13] *Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X. Y.* (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. *In Proceedings of the fourth ACM international conference on AI in finance* (p. 349–356). https://doi.org/10.1145/3604237.362686

[14] *Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E, Quoc V. Le, & Zhou D.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

[15] *Guo, Z., Cheng, S., Wang, Y., Li, P., & Liu, Y.* (2023). Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. https://doi.org/10.48550/arXiv.2305.17653

[16] *Hamotskyi, S., Levbarg, A. I., & Hänig, C.* (2024). Eval-UA-tion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models. *In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING* (p. 109–119). https://aclanthology.org/2024.unlp-1.13

[17] *Boroş, T., Chivereanu, R., Dumitrescu, S., & Purcaru, O.* (2024). Fine-Tuning and Retrieval Augmented Generation for Question Answering Using Affordable Large Language Models. *In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING* (p. 75–82). https://aclanthology.org/2024.unlp-1.10

[18] *Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., & Yin, R.* (2024). Searching for best practices in retrieval-augmented generation. https://doi.org/10.48550/arXiv.2407.01219

[19] *Taipalus, T.* (2024). Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research*, 85(7), 101216–101224. https://doi.org/10.1016/j.cogsys.2024.101216

ПОРІВНЯННЯ ZERO-SHOT ПІДХОДУ ТА ПОШУКУ З ДОПОВНЕНОЮ ГЕНЕРАЦІЄЮ ДЛЯ АНАЛІЗУ ТОНАЛЬНОСТІ КОМЕНТАРІВ УКРАЇНСЬКОЮ МОВОЮ

М. Притула, О. Сінькевич, І. Оленич

Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 м. Львів, Україна
marianna.prytula@lnu.edu.ua , *oleh.sinkevych@lnu.edu.ua* , *igor.olenych@lnu.edu.ua*

Вступ. Постійне зростання обсягів інформації, онлайн-новин і текстових повідомлень у соціальних мережах зумовлює нові виклики для суспільства. Розуміння емоційного аспекту безпосередньо впливає на задоволеність клієнтів у різних сферах діяльності та може вказувати напрямки для вдосконалення процесів. Тому, розробка засобів для аналізу тональності текстів є необхідною для забезпечення здатності якісного розпізнавання емоцій людей, виявлення проблем та визначення шляхів їх вирішення.

Методи. У роботі реалізовано підходи застосування великої мовної моделі Mistral-7B-UK для аналізу коментарів українською мовою. Використано два набори даних: один для бінарної класифікації, який розділено на негативний та позитивний класи, а інший – для багатокласової класифікації, що включав нейтральну тональність. Ці датасети містять відгуки про магазини, ресторани, готелі, медичні заклади, розважальні центри, фітнес-клуби, надання різноманітних послуг, тощо.

Результати. Розроблено підказки для zero-shot підходу, що включали роль, формат відповіді та додаткові пояснення щодо визначення тональності. Для впровадження пошуку з доповненою генерацією, використано векторну базу даних Qdrant та бібліотеку LangChain, яка забезпечила інтеграцію великої мовної моделі з зовнішніми джерелами даних. Під час визначення тональності тексту, п’ять найбільш семантично схожих фрагментів із визначеною тональністю повертаються з векторної бази даних і заповнюють попередньо визначені місця у шаблоні підказки. Далі відповідь моделі генерується з використанням наданого контексту. Проведено порівняння ефективності класифікації на основі отриманих результатів.

Висновки. Виявлено, що zero-shot підхід показує вищу точність для аналізу тональності коментарів порівняно з використанням пошуку з доповненою генерацією. Для багатокласової класифікації загальна точність становить 75 %, а для бінарної – 94 %. Використання зовнішніх ресурсів дозволило RAG моделі краще розпізнавати коментарі нейтральної тональності. Однак під час аналізу відповідей було виявлено, що коментарі протилежної тональності можуть бути отримані як контекст через спільний об’єкт опису, що в свою чергу негативно впливає на результати класифікації.

Ключові слова: тональність тексту, Велика мовна модель, zero-shot, пошук з доповненою генерацією.

The article was received by the editorial office on 18.06.2024.
Accepted for publication on 01.07.2024.