UDC 004.8

V. Oliinyk, N. Zakharchyn

A COMPARATIVE STUDY OF TASK FORMULATIONS FOR DETECTING PROPAGANDA USING LARGE LANGUAGE MODELS

Abstract: This paper extends existing studies on propaganda detection using large language models by examining several approaches to task formulation and applying them on different LLMs, namely, GPT-4o mini and Gemma / Gemma 2, aiming to find the most effective approach.

Using a combination of two text corpora in English and Russian languages with 18 propaganda techniques, we fine-tune models on character-based, phrase-based and class-?fication-only variations of this dataset with corresponding instructions to define which instruction yields the best performance. We conducted experiments and evaluated performance across classification, span identification, and joint tasks, demonstrating the clear superiority of the phrase-based approach over the character-based one. At the same time, our findings indicate that fine-tuning significantly improved model performance on span identification and joint tasks, while offering limited benefit for the classification task alone.

Keywords: propaganda detection, Large Language Models, propaganda techniques, fine-tuning, natural language processing.

Introduction

Propaganda detection is one of the most challenging and at the same time significant tasks in times when many political and social activities are taking place online. Propaganda can and is supposed to influence public opinions, manipulate people’s brains and distort reality. Therefore, the efforts to find effective ways to identify propaganda in news have been made not only in traditional fields like political sciences and journalism, but also by means of technologies, in particular, artificial intelligence.

The introduction of transformer architectures such as BERT, GPT and their advanced versions, large language models (LLMs), has opened up new horizons in natural language understanding. Due to their ability to process textual information and grasp the context, it is sensible to test their effectiveness on propaganda detection in news texts, which might bring solutions that can adapt to different languages and contexts.

As any other task in machine learning, propaganda detection task must be formalized. This is quite challenging, since there are many ways to define the meaning of the term “propaganda”. For instance, simple binary classification performed by the model on whether the text is propaganda or not might be insufficient because of the lack of traceability and reasoning in the answer. In this study, we explore identifying different techniques of propaganda and specific fragments that contain those techniques in the text, which is formally characterized as multi-class classification and span identification task.

We will study and compare performance of two families of large language models (LLMs), in particular, GPT-4o mini and Gemma / Gemma 2 on the same propaganda detection task, adopting different approaches to the task formulation and dataset format. In the process, we will use fine-tuning to calibrate models' responding style and combine it with few-shot learning to provide models with some examples.

We will use a dataset with both English and Russian languages, which contains articles and short Telegram posts and then compare performance of LLMs on those two languages.

Related works

The use of machine learning techniques for detecting manipulation, bias, and propaganda in media content has been explored in several studies. In [1], the authors address the challenge of propaganda detection in text by leveraging BERT models augmented with classifier layers for each level of granularity – text, paragraph, sentence, etc. These models aim to both classify the type of propaganda and identify the precise start-end character spans where it appears.

Similar formulations using BERT and other neural networks are found in several other works. For instance, [2] investigates various architectures including CNN, LSTM-CRF, and BERT for propaganda detection at both the sentence and fragment levels. Their approach integrates linguistic, layout, and thematic features, and employs multifaceted and multitask neural architectures. In addition, ensemble techniques such as majority voting and relaxation voting are applied to enhance performance.

The authors of [3] explore multi-class classification of propaganda techniques by fine-tuning large language models (LLMs), particularly from the GPT family. Their approach incorporates a "chain-of-thought" instruction, requiring models to provide reasoning for their classifications. Based on this idea, [4] presents a bilingual (English-Arabic) dataset of LLM-generated and manually evaluated explanations for fine-tuning, achieving competitive results with a fine-tuned LLaMA 3.1 model. Both studies focus on classification with justification, rather than direct identification of propaganda spans.

Finally, [5] highlights the limitations of LLMs in propaganda detection, emphasizing their unpredictability and lack of reliability for span-based tasks. This study also utilizes the character-span identification method for detecting propaganda.

Some studies also explore propaganda detection beyond news texts, focusing instead on social media platforms like Twitter (X). In [6], the authors introduce a dataset of weakly annotated tweets featuring fine-grained propaganda techniques and propose a neural approach for classifying tweets accordingly.

Our study contributes to this growing field by evaluating the effectiveness of LLMs in both propaganda identification and classification tasks. We focus particularly on how task formulation and dataset format affect performance. In addition to the classification task, we explored two approaches to span identification: detecting start- and end-character spans of propaganda, and identifying propaganda-containing phrases. To our knowledge, this phrase-based approach has not yet been investigated for propaganda detection in combination with large language models.

Fine-tuning dataset

The dataset used for fine-tuning the LLMs has the following structure: each record consists of a news piece (either a news article or a Telegram post), the identified propaganda techniques, and the corresponding character spans in the text where those techniques appear.

It combines two sources with identical annotation formats. The first is from [1], comprising 451 English-language news articles from 48 different sources, annotated with 18 propaganda techniques. The second source is a corpus from Mantis Analytics [7], containing 1,357 Telegram news posts in Russian, annotated using the same methodology.

Together, the merged dataset includes 1,700 samples, split into training (1,172), validation (348), and inference (180) sets. The dataset is imbalanced, as some propaganda techniques are infrequently represented in the texts. Sample lengths vary significantly—from 10 to 47,706 characters. The class distribution is illustrated in Fig. 1.

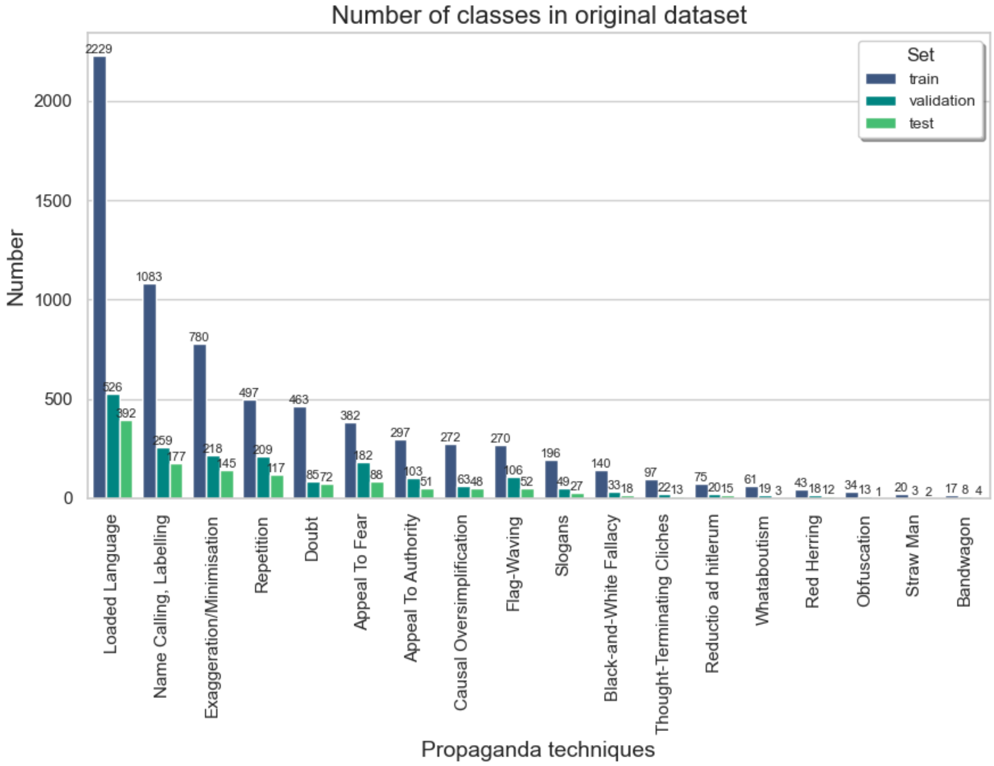


Figure 1. Class Distribution Histogram

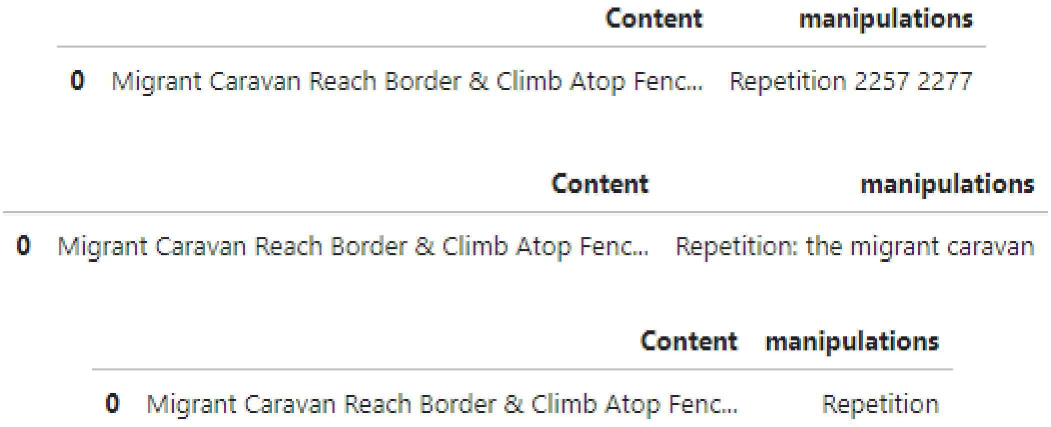
The initial dataset was preprocessed by removing duplicates, eliminating unnecessary characters, and standardizing the format of the column containing propaganda techniques. To evaluate the performance of LLMs under different conditions, we created several dataset variations for fine-tuning:

– Original Format. Contains a Content column and a Manipulations column, with each entry listing propaganda techniques along with their corresponding start- and end-character spans. This format supports a joint task of multi-class classification and span identification.

– Original format with propaganda techniques and a fragment from the text containing each technique. instead of solely relying on character span annotations (as in Variation 1), we augmented the "Manipulations" column to include full phrases containing propaganda. This modification allowed us to directly compare the performance of Large Language Models (LLMs) when identifying numerical spans versus identifying textual instances of propaganda.

– Classification Format. Includes a Content column and a Manipulations column listing only the propaganda techniques, focusing solely on the classification task.

Representative examples from all three dataset formats are presented in Fig. 2.



	Content	manipulations
0	Migrant Caravan Reach Border & Climb Atop Fenc...	Repetition 2257 2277

	Content	manipulations
0	Migrant Caravan Reach Border & Climb Atop Fenc...	Repetition: the migrant caravan

	Content	manipulations
0	Migrant Caravan Reach Border & Climb Atop Fenc...	Repetition

Figure 2. Examples from dataset variations

Experimental setup

The large language models used in this study include:

– **GPT-4o Mini.** The most cost-effective and accessible model from OpenAI [8] at the time of experiment. Both fine-tuning and inference are performed via the OpenAI API, eliminating the need for local computational resources.

– **Gemma 7B** and **Gemma 2 9B** from Google Research [9]: Among the largest parameter models available for local deployment on GPU/TPU hardware.

We conducted experiments fine-tuning these LLMs on the dataset variations described earlier. For the joint classification and span identification task, we fine-tuned GPT-4o Mini and Gemma 2 9B — the two largest models in terms of available resources. For the classification-only task, all models were fine-tuned. Table 1 summarizes the combinations of LLMs and dataset variations used in the experiments.

Table 1.

LLMs and datasets used for experiments

	GPT-4o mini, fine-tuned	**Gemma 7B, fine-tuned**	**Gemma 2 9B, fine-tuned**
Classification dataset	Yes	Yes	Yes
Dataset with character spans	Yes	No	Yes
Dataset with phrases	Yes	No	Yes

For fine-tuning and inference, we adopted a base prompt from [3], experimenting with different formatting styles, shortening certain parts, and modifying the instruction section depending on the specific task variation.

Although OpenAI provides metrics like accuracy and loss during training and validation for GPT-4o Mini, these metrics do not fully capture real-world performance. Therefore, we use the F1 score as the primary evaluation metric, both averaged overall and per class.

For the span identification task, a predicted phrase is considered a positive match if its similarity to the actual phrase exceeds a defined threshold. This threshold accounts for cases where the predicted fragment is slightly longer or shorter than the true span by a few neighboring words. After testing different values, we set the similarity threshold at 0.5.

Metrics are computed separately for the classification task, the span identification task (where applicable), and the combined joint task. This approach allows us to evaluate which task the LLM performs better on and by what margin. Additionally, we calculate metrics for the baseline GPT-4o Mini model without fine-tuning to assess the impact of fine-tuning.

Finally, inference results will be analyzed separately for the English and Russian subsets of the dataset.

Experimental Results

We began by running inference on the GPT-4o mini models to determine which task formulation and dataset variation led to the best performance. Results showed that GPT-4o mini performed best when fine-tuned and evaluated on a dataset containing direct propaganda phrases, rather than character span annotations.

Accordingly, we used this setup as the baseline to compare against the best fine-tuned version. F1 score results for different GPT-4o mini configurations are presented in Table 2.

Table 2.

F1 scores for GPT-4o mini models

	GPT-4o mini, dataset with phrases	**GPT-4o mini, dataset with spans**	**GPT-4o mini, classification dataset**	**GPT-4o mini, baseline**
Classification	0.45	0.45	0.45	0.45
Span identification	0.47	0.47	0.47	0.47
Joint task	0.48	0.48	0.48	0.48

It is evident that the phrase-based approach yields the best results for propaganda span identification and, consequently, for the joint task as well. In contrast, the character-based approach performs significantly worse. We attribute this to the fact that large language models (LLMs) generally struggle with tasks involving character-level calculations. Due to tokenization mechanics, an LLM may interpret a single word as multiple tokens, which can disrupt accurate character counting – an issue highlighted in experiments such as the Strawberry Test [10]. Since LLMs are primarily designed for natural language understanding and generation, they are not inherently optimized for tasks requiring low-level text manipulation. However, as noted in the Related Works section, character-based span identification can still be effective when LLMs are integrated into larger systems, particularly when supported by additional neural network layers.

One potential issue that we were expecting to arise with the LLM fine-tuned on phrases dataset were LLM hallucinations: models responding with phrases that were not presented in the original text. Although we did not track this issue formally, it was not encountered during manual assessment of model outputs.

Regarding the classification task, the model fine-tuned specifically on the classification dataset variation demonstrated an insignificant performance advantage over the other fine-tuned models. However, fine-tuning did not significantly enhance classification performance overall. Interestingly, the baseline GPT-4o mini outperformed all fine-tuned variants in this task. Nonetheless, we observed clear improvements from fine-tuning in both the span identification and joint tasks, underlining the effectiveness of task-specific tuning for more complex use cases.

We then conducted the same experiments on Gemma / Gemma 2 models. Results are shown in Table 3.

We again observe the tendency that Gemma 2 fine-tuned model performs worse on the character-based task than on phrase-based, strengthening our hypothesis that phrase-based approach to the task formulation is a better fit for LLMs. We also conclude that tested variations are inferior to GPT-4o mini when it comes to propaganda detection in the given problem formulation. Consequently, we focused our further experiments on better-performing models.

Table 3.

F1 scores for Gemma models

	Gemma 2 9B, dataset with phrases	**Gemma 2 9B, dataset with spans**	**Gemma 2 9B, classification dataset**	**Gemma 7B, classification dataset**
Classification	0.30	0.30	0.30	0.30
Span identification	0.30	0.30	0.30	0.30
Joint task	0.45	0.45	0.45	0.45

Using the LLM variation with the highest score for the joint task (GPT-4o mini, fine-tuned on the dataset with phrases), we reviewed F1 scores for each propaganda technique class. Results for GPT-4o mini fine-tuned on phrases dataset are depicted in Table 4.

Table 4.

F1 scores per class

Class	**F1 score**
Appeal_to_Authority	0.407
Appeal_to_fear-prejudice	0.32
Bandwagon	0.0
Black-and-White_Fallacy	0.143
Causal_Oversimplification	0.118
Doubt	0.4
Exaggeration,Minimisation	0.511
Flag-Waving	0.292
Loaded Language	0.789
Name Calling, Labeling	0.554
Obfuscation,Intentional_Vagueness,Confusion	0.0
Red_Herring	0.0
Reductio_ad_hitlerum	0.333
Repetition	0.205
Slogans	0.191
Straw_Man	0.0
Thought-terminating_Cliches	0.0
Whataboutism	0.0

The results clearly indicate that class imbalance in the original dataset significantly affects model performance, as evidenced by zero F1 scores for the least represented classes. In contrast, the most accurately detected technique was “Loaded Language”, which aligns with its high frequency in the training data. Increasing the number of samples for underrepresented classes could potentially enhance the model's overall performance.

Table 5.

F1 scores per language

	Classification	Span identification	Joint task
English	0.48	0.48	0.48
Russian	0.48	0.48	0.48

Finally, using the best-performing model, we computed F1 scores separately for the English and Russian subsets of the dataset. The comparison is presented in Table 5.

As expected, the model performs better on English articles, likely due to the fact that Russian is a more token-expensive language. To further increase the score for Russian language as well as for any low-resource languages, it might be beneficial to explore training data augmentation techniques, such as machine translation, as demonstrated in [11] and [12].

Comparison with previous work

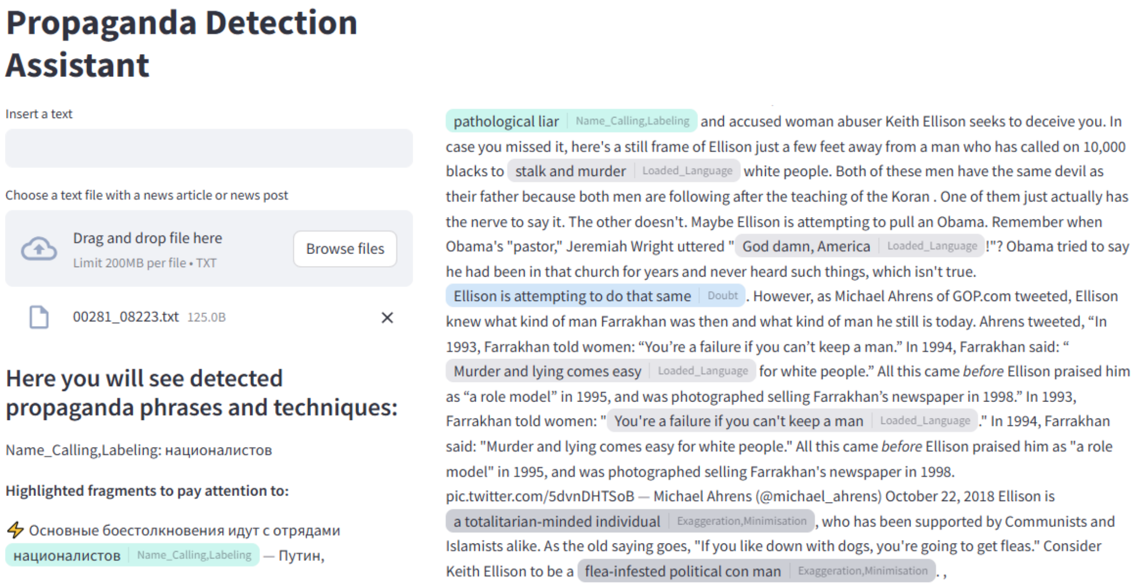
The authors of [1], although they also calculated F1 scores for span identification and joint tasks, did not appear to use a similarity threshold approach. Therefore, while our results are numerically higher (F1 score of 44% versus 23%), a direct comparison is not entirely objective.

However, we can compare our classification results with those presented in [3]. The authors reported an F1 score of 58% for the classification task using a fine-tuned GPT-4 model. Our highest classification score reaches 48%, with similar class-wise trends—particularly, some classes receiving an F1 score of 0. The difference in overall performance may be attributed to the use of different language models, as GPT-4 has a significantly larger parameter size. Additionally, the referenced study used only an English-language corpus, which may have further contributed to their higher results. Notably, their reported F1 score for GPT-3 was 44%, which our model surpasses.

The work [4] presents F1 scores for their binary classification approach focused on explanation-based propaganda detection. While their primary focus is on evaluating the quality of LLM-generated explanations, a common conclusion drawn from both their work and ours is that LLMs perform better on English compared to lower-resourced languages such as Arabic and Russian.

Practical application

To showcase and evaluate the results of our propaganda detection models, we developed a web-based application that provides a user-friendly interface for viewing detected fragments and associated techniques. This tool allows for intuitive interaction with the model outputs and is intended for both demonstration and analysis purposes. Fig. 3 illustrates example use cases of the application.


Propaganda Detection Assistant
Insert a text
Choose a text file with a news article or news post
Drag and drop file here
Limit 200MB per file • TXT
Browse files
00281_08223.txt 125.0B
×

Here you will see detected propaganda phrases and techniques:

Name_Calling,Labeling: националистов

Highlighted fragments to pay attention to:

⚡ Основные боестолкновения идут с отрядами националистов Name_Calling,Labeling — Путин,

pathological liar Name_Calling,Labeling and accused woman abuser Keith Ellison seeks to deceive you. In case you missed it, here's a still frame of Ellison just a few feet away from a man who has called on 10,000 blacks to stalk and murder Loaded_Language white people. Both of these men have the same devil as their father because both men are following after the teaching of the Koran . One of them just actually has the nerve to say it. The other doesn't. Maybe Ellison is attempting to pull an Obama. Remember when Obama's "pastor," Jeremiah Wright uttered " God damn, America Loaded_Language !"? Obama tried to say he had been in that church for years and never heard such things, which isn't true. Ellison is attempting to do that same Doubt . However, as Michael Ahrens of GOP.com tweeted, Ellison knew what kind of man Farrakhan was then and what kind of man he still is today. Ahrens tweeted, “In 1993, Farrakhan told women: “You’re a failure if you can’t keep a man.” In 1994, Farrakhan said: “ Murder and lying comes easy Loaded_Language for white people.” All this came *before* Ellison praised him as “a role model” in 1995, and was photographed selling Farrakhan’s newspaper in 1998.” In 1993, Farrakhan told women: " You're a failure if you can't keep a man Loaded_Language ." In 1994, Farrakhan said: "Murder and lying comes easy for white people." All this came *before* Ellison praised him as "a role model" in 1995, and was photographed selling Farrakhan's newspaper in 1998. pic.twitter.com/5dvnDHTSoB — Michael Ahrens (@michael_ahrens) October 22, 2018 Ellison is a totalitarian-minded individual Exaggeration,Minimisation , who has been supported by Communists and Islamists alike. As the old saying goes, "If you like down with dogs, you're going to get fleas." Consider Keith Ellison to be a flea-infested political con man Exaggeration,Minimisation . ,

Figure 3. Examples of propaganda detection results displayed in the web-based application

Conclusion

In this paper, we explored the use of large language models, specifically GPT-4o mini and Gemma / Gemma 2, for the task of propaganda detection.

We examined how different task formulations and fine-tuning configurations impact model performance. Three variations were tested: (1) identifying start- and end-character spans of propaganda-containing fragments along with their associated techniques, (2) identifying text fragments containing propaganda with their corresponding techniques, and (3) identifying only the propaganda techniques present. These variations were evaluated across three tasks: technique classification, span (fragment) identification, and a combined joint task.

Our results demonstrate that phrase-based task formulations outperform character-based ones in both span identification and the joint task, achieving higher F1 scores. This supports the observation that LLMs struggle with character-level numerical reasoning due to tokenization limitations. Accordingly, we recommend using phrase-based approaches for more effective detection of propaganda fragments.

Finally, we found that model performance was notably better on English texts than on Russian ones, a likely consequence of the latter being a more under-resourced language.

It is worth noting that detecting propaganda in news is a complex task that goes beyond simple classification. It often demands critical thinking, contextual understanding, and background knowledge – elements that may be lacking in the limited text input provided to LLMs. While LLMs are not yet perfect for this task, they can still be effectively used to highlight potentially problematic parts of a text and guide users towards deeper analysis.

Our future work will be aimed at overcoming challenges of imbalanced data, exploring the use of other LLMs as well as experimenting with task formalization.

References

1. Da San Martino, G., Seunghak, Y., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019). Fine-grained Analysis of Propaganda in News Articles. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 5636-5646). Association for Computational Linguistics.

2. Gupta, P., Saxena, K., Yaseen, U., Runkler, T., & Schütze, H. "Neural architectures for fine-grained propaganda detection in news." arXiv preprint arXiv:1909.06162, 2019.

3. Sprenkamp, Kilian, Daniel Gordon Jones, Liudmila Zavolokina. "Large language models for propaganda detection." arXiv preprint arXiv:2310.06422, 2023.

4. Hasanain, M., Hasan, M. A., Kmainasi, M. B., Sartori, E., Shahroor, A. E., Martino, G. D. S., & Alam, F. "Reasoning About Persuasion: Can LLMs Enable Explainable Propaganda Detection?". arXiv preprint arXiv:2502.16550, 2025.

5. Szwoch, J., Staszkow, M., Rzepka, R., & Araki, K. Limitations of large language models in propaganda detection task. Applied Sciences, 14(10), 4330, 2024.

6. Vijayaraghavan, Prashanth, and Soroush Vosoughi. "TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.

7. Disinformation Detection Challenge by AI HOUSE x Mantis Analytics. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/competitions/disinformation-detection-challenge/data (application date 28.04.2025)

8. GPT-4o mini: advancing cost-efficient intelligence. OpenAI. URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence (application date: 28.04.2025)

9. Gemma: Open Models Based on Gemini. Research and Technology. URL: https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf (application date 28.04.2025)

10. Why LLMs Can't Count the R's in 'Strawberry' & What It Teaches Us. URL: https://arbisoft.com/blogs/why-ll-ms-can-t-count-the-r-s-in-strawberry-and-what-it-teaches-us#the-case-of-strawberry (application date: 28.04.2025)

11. Oliinyk V. Data augmentation with foreign language content in text classification using machine learning / Oliinyk V., Osadcha K. // Adaptive systems of automatic control, 2020. Vol. 1, №36. – P. 51-59.

12. Oliinyk V. Low-resource text classification using cross-lingual models for bullying detection in the Ukrainian language / Oliinyk V., Matviichuk I. // Adaptive systems of automatic control, 2023. Vol. 1, №42. – P. 87-100.