Fine-Tuned Transformers for Detection and Classification of Persuasion Techniques in Slavic Languages

Ekaterina Loginova
Oplot / -
ekaterina.d.loginova@gmail.com

Abstract

This paper details a system developed for the SlavicNLP 2025 Shared Task on the Detection and Classification of Persuasion Techniques in Texts for Slavic Languages. The shared task comprises two subtasks: binary detection of persuasive content within text fragments and multi-class, multi-label identification of specific persuasion techniques at the token level. Our primary approach for both subtasks involved fine-tuning pre-trained multilingual Transformer models. The resulting systems reached F1 score of 0.92 in paragraph-level detection (ranked third on average). We present our system architecture, data handling, training procedures, and official results, alongside areas for future improvement.

1 Introduction

Persuasion techniques, ranging from loaded language to false dilemmas, play a central role in propaganda and manipulation. Automatically identifying such techniques is therefore a critical step towards trustworthy media ecosystems. However, recent surveys highlight the scarcity of Slavic-language resources for such tasks. The SlavicNLP 2025 shared task (Piskorski et al., 2025) addresses this gap with two subtasks: binary detection of persuasive paragraphs (Subtask 1) and fine-grained multi-label span classification (Subtask 2).

Reliable persuasion detection is demanding due to the subtlety of persuasive language, the potential for multiple techniques co-occurring within a single fragment, and the inherent difficulty in distinguishing legitimate argumentation from manipulative rhetoric. Transformers (Vaswani et al., 2023) have been applied in recent work on propaganda and manipulation detection in multilingual settings (Solopova et al., 2024), and our system builds on this trend. For Subtask 1, we fine-tuned a Transformer for sequence classification to make binary predictions at the paragraph level. We enhanced this approach by incorporating a small corpus of additionally manually labelled data. For the more granular Subtask 2, we fine-tuned Transformer models for token classification. The following sections outline our system implementation, performance analysis and methodological insights.

2 Data

The shared task included texts from parliamentary debates and social media posts across five Slavic languages, employing an extended version of the SemEval 2023 Task 3 persuasion technique taxonomy, which includes 25 fine-grained techniques across 6 main categories. Each instance is thus a paragraph with one or more persuasion spans annotated using the taxonomy. A notable characteristic of the dataset was the imbalance in label distribution, with techniques such as 'Loaded Language', 'Name Calling / Labeling', and 'Repetition' being significantly more prevalent than others like 'Appeal to Pity'.

For Subtask 1, we used SemEval 2023 data (Piskorski et al., 2023) and augmented the dataset with additional 2821 sentences from 260 texts from Russian state-sponsored and opposition news channels, annotated in-house by four volunteers. Overall inter-annotator agreement (average pairwise Cohen's Kappa) was 0.60, Fleiss' Kappa was 0.62. Cosine similarity of means with the shared task's dataset was 0.85, using 'all-MiniLM-L6-v2' Sentence Transformers model and Wasserstein Distance on first PCA dimension was 0.38, indicating an extension of the original domain. The dataset is available at request.

Technical details. Training subset included the shared task's training data and SemEval 2023 data, validation subset included the trial part of the shared task's data. No additional preprocessing was done. In the Subtask 2, labels were aligned with the tokenized output using the IOB tagging

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 151–156

scheme.

3 Models

Our submitted systems for both subtasks were fine-tuned multilingual Transformer models. We have also trained traditional machine learning models as baselines and experimented with LLM prompt engineering.

3.1 Subtask 1.

To determine the presence or absence of any persuasion technique within a given text fragment, we implemented a binary sequence classification approach. As a baseline, we implemented SVM (Cortes and Vapnik, 1995) and XGBoost (Chen and Guestrin, 2016) on TF-IDF.

For fine-tuning, we have considered two strategies. The most straightforward approach is to label a binary classifier on the target label. A more specific strategy involved training 25 distinct binary classifiers, one for each persuasion technique. The final label was then inferred if at least one of the individual classifiers yielded a positive prediction. Both strategies can be extended by training language-specific models; however, we opted not to pursue this direction, due to the limited number of training samples per language. Our primary approach was a single multilingual model. As base models, we experimented with FacebookAI/xlm-roberta-base (Conneau et al., 2020), sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019), google-bert/bert-base-multilingual-cased (Devlin et al., 2018), intfloat/multilingual-e5-small (Wang et al., 2024). Based on the performance comparison on the validation set, we initially chose intfloat/multilingual-e5-small and that was the base models for the solution submitted within the deadline, but after correction to the data selection code, we switched to using FacebookAI/xlm-roberta-base (see 1). As a result of data selection adjustment, we corrected the drop in performance for Russian language that was observed on the official test set. See comparison of results in Table 4.

Models were trained using the Hugging Face Trainer API. Learning rate was set in the range of 2e-5 to 5e-5, batch sizes of 16 or 32, and training for 3 to 5 epochs. An early stopping was used to prevent overfitting. Model training was conducted using Google Colaboratory, NVIDIA T4 GPUs.

3.2 Subtask 2.

For the task of identifying the exact spans and types of specific persuasion techniques, we adopted a token classification framework, using the same base models as in Subtask 1. Concerning the training regimen: learning rates of 2e-5, a batch size of 4 (due to the higher memory demands of token classification), and training for up to 15 epochs. Weight decay (e.g., 0.01) was applied as a regularization technique. AdamW (Loshchilov and Hutter, 2019) optimizer was used, with a linear learning rate decay schedule.

4 Results

4.1 Subtask 1.

The model generalises well to unseen languages (see Table 2). The surprising result was low performance for Russian, which we attribute to a data selection error, which we corrected after the official competition deadline (see the comparison of results in Table 4). Automated label quality checks by cleanlab[1] library also suggested that potential label noise or inconsistencies in the training data might have impacted model learning (32% of the data affected).

4.2 Subtask 2.

The token classification approach for identifying specific techniques proved challenging. Initial experiments with google-bert/bert-base-multilingual-cased and FacebookAI/xlm-roberta-base on subsets of the data yielded overall F1 scores (micro-averaged across all technique classes) in the lower range (e.g., 0.02 to 0.06). These preliminary figures underscore the difficulty of precise token-level multi-label classification across imbalanced classes. The final official test results were provided by the organisers (**?**). As one can see, token-level scores are considerably lower, owing to strict span boundaries and severe label imbalance.

5 Experiments

The following results were obtained after the official submission and are not part of our primary system.

5.1 Subtask 1.

Per-technique binary classifiers. The strategy of training individual binary classifiers per technique

[1] https://github.com/cleanlab/cleanlab

Model Type	Base Model	Extra Data	Strategy	F1 Score
Traditional ML	TF-IDF + SVM	No	Single full text classifier	0.53
Traditional ML	TF-IDF + XGBoost	No	Single full text classifier	0.59
Transformer	e5-small	No	Single full text classifier	0.79
Transformer	e5-small	Yes	Single full text classifier	0.76
Transformer	xlm-roberta-base	No	Single full text classifier	0.77
Transformer	xlm-roberta-base	Yes	Single full text classifier	0.82
Transformer	bert-base-multilingual-cased	No	Single full text classifier	0.80
Transformer	bert-base-multilingual-cased	Yes	Single full text classifier	0.80
Transformer	MiniLM-L12-v2	No	Single full text classifier	0.78
Transformer	MiniLM-L12-v2	Yes	Single full text classifier	0.76
LLM	Claude Sonnet 3.7	No	Zero-shot	0.74
LLM	Claude Sonnet 3.7	No	Few-shot	0.82
LLM	GPT-4o	No	Zero-shot	0.83
LLM	GPT-4o	No	Few-shot	0.83
LLM	Gemini 1.5 Pro	No	Zero-shot	0.79
LLM	Gemini 1.5 Pro	No	Few-shot	0.65

Table 1: Subtask 1. Validation set F1 scores of different model types, base models, and strategies. (Results after data selection correction.)

Language	F1 (test)
Bulgarian	0.87
Croatian	0.92
Polish	0.90
Russian	0.83
Slovene	0.85

Table 2: Subtask 1. Official evaluation F1 scores on the test set, by language.

Language	Macro F1	Micro F1
Bulgarian	0.1850	0.1983
Croatian	0.2772	0.2709
Polish	0.2111	0.2015
Russian	0.1289	0.2126
Slovene	0.1131	0.1786

Table 3: Subtask 2. Official evaluation macro and micro F1 scores on the test set, by language.

yielded varying F1 scores depending on the specific technique, the hardest to predict being 'Appeal to Pity' at 0.71 and the simplest being Consequential Oversimplification at 0.87, with the average of 0.81. However, combining the results achieved only 0.64 on the final prediction task, dependening on the probability threshold (0.95 appeared optimal).

Traditional machine learning models, while computationally efficient, generally underperformed compared to fine-tuned Transformers, yielding F1 scores in the 0.67-0.73 range. 5-fold stratified cross-validation with hyperparameter grid-search, using scikit-learn (Pedregosa et al., 2011) package. TF-IDF is the only scenario where we applied text preprocessing: filtering by stop-words and part of speech, then lemmatising using spaCy (Honnibal et al., 2020).

Language	Before	After
Overall	0.80	0.84
bg (Bulgarian)	0.75	0.80
pl (Polish)	0.90	0.90
ru (Russian)	0.49	0.72
si (Slovenian)	0.88	0.78

Table 4: Subtask 1 validation set F1 scores before and after data selection adjustment.

We have also evaluated proprietary LLMs, namely Anthropic Claude Sonnet 3.7, OpenAI GPT 4o and Google Gemini 1.5 Pro. As demonstrated in Table 1, traditional machine learning models performed modestly, with F1 scores ranging from 0.53 (TF-IDF + SVM) to 0.59 (TF-IDF + XGBoost). Transformer-based models showed strong performance, achieving up to 0.82 with 'xlm-roberta-base' and extra data, and stable scores around 0.76–0.80 across other multilingual models and settings. Among LLMs, OpenAI's GPT-4o achieved the highest F1 score (0.83) consistently in both zero-shot and few-shot setups. Claude Sonnet 3.7 and Gemini 1.5 Pro also performed well, though

Gemini showed a notable drop in few-shot prompting (0.65). Overall, LLMs outperformed traditional models, and few-shot prompting often provided gains, except in the case of Gemini.

For the few-shot setup, examples were chosen randomly. A possible modification is to find the semantically closest text to the one being evaluated. The LiteLLM Python package [2] was used to benefit from a uniform prompting interface. The prompts included the the hierarchical list of available manipulation techniques. We have also experimented with providing short explanations of each method and examples, based on the taxonomy description paper, but that did not improve the results. The temperature was set to 0; measuring the influence of this parameter is a prospective research question.

5.2 Subtask 2.

As can be seen in Table 5, all three LLMs (Claude Sonnet 3.7, GPT-4o, and Gemini 1.5 Pro) achieved near-identical micro F1 scores around 0.97–0.98 in zero-shot setup (likely due to the majority class of 'O'), but their macro scores remained much lower at 0.49, indicating uneven performance across classes. This suggests strong overall accuracy but challenges with class imbalance or underrepresented labels.

After the end of the test phase, we experimented with a two-step approach where the first token classification model detects spans that contain any persuasion techniques (so the classes are O, I-MANIPULATION, B-MANIPULATION). Then, the second model, multi-label classification, predicts the label for each span. Preliminary experiments show that this approach improves precision, but recall drops drastically. As such, we will continue the investigation.

LLM	F1 (micro / macro)
Claude Sonnet 3.7	0.97 / 0.49
GPT-4o	0.97 / 0.49
Gemini 1.5 Pro	0.98 / 0.49

Table 5: Micro and macro F1 scores of LLMs under zero-shot setup on the validation set.

6 Discussion

The detection and classification of persuasion techniques present a formidable challenge. Our fine-tuned Transformer-based systems achieved promising results, especially for the binary detection subtask. The token classification approach for fine-grained classification, while offering detailed localisation, faced greater hurdles due to task complexity and data characteristics.

The high Subtask 1 scores confirm that paragraph-level propaganda cues are well captured by multilingual Transformers. Conversely, the poor Subtask 2 performance can be attributed to: (i) extreme class imbalance; (ii) sparsity of token-level signal; (iii) subtle boundary definitions (Loaded Language vs. Name Calling), with models struggling to predict less frequent persuasion techniques.

Using fine-tuned Transformer models instead of large generative LLMs like GPT-4 for the SlavicNLP 2025 shared task—especially in detecting and classifying propaganda techniques—offers several practical and methodological advantages. LLMs like GPT-4 are generative, making them less reliable for consistent classification, especially for span-level tasks, where subtle variations can lead to inconsistent labels. Transformer classifiers, on the contrary, offer repeatable predictions, which is essential for creating transparent and auditable models, especially in sensitive domains like media manipulation detection.

Fine-tuned Transformers (base or distilled) can be efficiently deployed on modest hardware, supporting large-scale processing needs. In many practical applications (e.g., media monitoring systems), the goal is to aggregate manipulation indicators across sources or time, and flag patterns or early warnings of coordinated propaganda. LLMs are less suited to this, as their cost limits scalability. Choosing simpler models is both more environmentally friendly and more accessible for independent media outlets, NGOs and activists.

Furthermore, LLM inference often requires API calls to external servers, introducing privacy concerns and dependency on proprietary infrastructure. LLMs often depend on cloud access (e.g., OpenAI API), making them unsuitable for privacy-sensitive or legally regulated contexts (like NGO deployments in authoritarian environments). For sensitive data (e.g., monitoring fringe political channels, Telegram groups), local deployment is a must.

Recent academic research increasingly demonstrates that Large Language Models (LLMs), despite often being presented as neutral information processors, can exhibit strong political biases (Peng et al., 2025). These biases can manifest in various

[2] https://www.litellm.ai/

ways, from favoring certain political ideologies and figures to framing information in a skewed manner, potentially influencing user perception and public discourse. Propaganda detection is politically sensitive, whereas LLMs might reflect systemic biases, especially when applied to Slavic languages or contentious geopolitical contexts.

Nevertheless, LLMs offer impressive generalisation in few-shot setups, and for evaluating implicit bias, rhetorical coherence, or generating explanations, they shine. Therefore, for end-to-end user-facing applications, we recommend using the best of both worlds: Transformers for fast detection, aggregated statistics and preliminary analysis, and LLMs for human-facing explanations or validation.

In the future, we plan to conduct a thorough manual review of model errors on a validation set to gain deeper insights into misclassification patterns across different techniques and languages, and leverage LLMs more extensively for targeted data augmentation, particularly for underrepresented techniques and complex cases. For practical applications, such as tools for media literacy or content moderation aids, it is crucial to evaluate models for potential biases learned from the training data. Such biases could lead to disproportionate flagging of content from certain demographic groups or a failure to detect manipulation targeted at specific communities. Regular audits for fairness across languages, topics, and author demographics would be necessary. Persuasion tactics evolve; deployed systems will require ongoing monitoring and periodic retraining with new data to maintain their effectiveness against emerging techniques.

Acknowledgements

We would like to thank the volunteers, Maryia Marynich and others, who supported our project by annotating additional training data.

7 Conclusion

This paper has outlined our approach to the SlavicNLP 2025 Shared Task, centered on the application of fine-tuned multilingual Transformer models for detecting and classifying persuasion techniques. Our findings indicate that while Transformers are potent tools for these tasks, challenges related to data imbalance, linguistic nuance, and the inherent complexity of persuasion persist. The binary detection task yielded more robust results, while the fine-grained token-level classification proved more demanding. Future efforts should concentrate on sophisticated data augmentation, advanced model architectures, and comprehensive error analysis to advance the capabilities of automated persuasion technique identification in diverse linguistic contexts. The shared task has provided valuable insights into the intricacies of modelling persuasive language and the ongoing need for research in this critical area.

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tai-Quan Peng, Kaiqi Yang, Sanguk Lee, Hang Li, Yucheng Chu, Yuping Lin, and Hui Liu. 2025. Beyond partisan leaning: A comparative analysis of political bias in large language models.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobrani'c, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubeši'c, Michał Marci'nczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. 2025. SlavicNLP 2025 Shared Task: Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing 2025 (SlavicNLP 2025)*, Vienna, Austria. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Veronika Solopova, Viktoriia Herman, Christoph Benzmüller, and Tim Landgraf. 2024. Check news in one click: Nlp-empowered pro-kremlin propaganda detection. *arXiv preprint arXiv:2401.15717*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

A Appendix

Few-Shot Prompt for Subtask 1 (Binary Text Classification)

```
"You are an expert linguist.whether the following text employs any of the persuasion techniques listed below.only one line of JSON exactly in the format:{ "propaganda": 0 or 1}:{TAXONOMY}{examples}: "{text}":
```

Few-Shot Prompt for Subtask 2 (Multi-label Token Classification)

```
"You are a linguist detecting manipulation in news texts. Your task is to find spans of text that match any of the following persuasion techniques, and assign a technique label to each span. Output must be a JSON list of objects, each with: ‘start‘ (char index), ‘end‘ (exclusive), ‘technique‘ (from the list below):{TAXONOMY}:{text}:"
```