Gradient Flush at Slavic NLP 2025 Task: Leveraging Slavic BERT and Translation for Persuasion Techniques Classification

Sergey Senichev, Aleksandr Boriskin, Nikita Krayko, Daria Galimzianova
MTS AI

Abstract

The task of persuasion techniques detection is limited by several challenges, such as insufficient training data and ambiguity in labels. In this paper, we describe a solution for the Slavic NLP 2025 Shared Task. It utilizes multilingual XLM-RoBERTa, that was trained on 100 various languages, and Slavic BERT, a model fine-tuned on four languages of the Slavic group. We suggest to augment the training dataset with related data from previous shared tasks, as well as some automatic translations from English and German. The resulting solutions are ranked among the top 3 for Russian in the Subtask 1 and for all languages in the Subtask 2. We release the code for our solution.[1]

1 Introduction

The increasing prevalence of persuasive techniques in political discourse and social media has raised concerns about their role in shaping public opinion, particularly in the context of disinformation and manipulative communication. Detecting and classifying such techniques in multilingual settings presents a significant challenge due to linguistic and cultural variations, especially in under-resourced languages.

The Shared Task on the Detection and Classification of Persuasion Techniques in Texts for Slavic Languages (Piskorski et al., 2025) addresses this challenge by focusing on five Slavic languages: Bulgarian, Polish, Croatian, Slovene, and Russian across two domains, parliamentary debates and social media posts. The task consists of two subtasks: (1) binary detection of persuasion techniques in text fragments and (2) multi-label classification of specific techniques based on an extended taxonomy derived from SemEval 2023 Task 3 (Piskorski et al., 2023a).

In this paper, we propose a solution leveraging multilingual transformer-based models, specifically XLM-RoBERTa (Conneau et al., 2019) and Slavic BERT (Arkhipov et al., 2019), which are well-suited for cross-lingual transfer learning. Given the limited labeled data for some languages, we employ training data augmentation by incorporating official datasets alongside translated examples from related persuasion technique corpora. This approach enhances model generalization across languages while mitigating data scarcity.

For Subtask 1, which involves binary classification of whether a given text offset contains any persuasion techniques, our Slavic BERT-based solution *is ranked 2nd for the Russian language*. Subtask 2 is a multi-label classification problem: given a text offset and a predetermined list of persuasion techniques, determine which of the techniques are present in the text. For this task, our approach with XLM-RoBERTa and augmentation from other related sources *is ranked among the top 3 for three out of five languages* measured with both micro and macro F1 scores. For the remaining two languages (Bulgarian and Russian), this method made it to the top 3 on one of the scores.

2 Related work

In recent years, multilingual persuasion techniques detection (Martino et al., 2020), which was inspired by disinformation research and the analysis of political and social-media rhetoric, has gained attention. SlavicNLP-2025 shared task concentrates these efforts specifically on five Slavic languages: Bulgarian, Polish, Croatian, Slovene and Russian.

Various shared tasks related to the detection of Persuasion Techniques, moving from news (SemEval-2023 Task 3) (Piskorski et al., 2023a) to social media (SemEval-2020 Task 11) (Da San Martino et al., 2020) and even multimodal memes (SemEval-2021 Task 6) (Dimitrov et al.,

[1] https://github.com/ssenichev/ACL_SlavicNLP2025/

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 171–176

2021).

(Da San Martino et al., 2020) started the research with SemEval-2020 Task 11, which treated persuasion as *span identification* and *span classification* in news articles, using an inventory of 14 techniques. The research quickly shifted to multimodality: Dimitrov et al. (2021) organized SemEval-2021 Task 6, releasing a 950-meme corpus annotated with 22 techniques and defining three subtasks that jointly exploit text and image cues. Building on previous tasks, SemEval-2023 Task 3 (Piskorski et al., 2023a) released a multilingual news corpus in nine languages (including PL and RU), annotated with 23 persuasion techniques at the paragraph level. Most recently, "CLEF-2024 CheckThat! Task 3" (Piskorski et al., 2024) has also included 4 Slavic languages (RU, BG, PL, SL) in their datasets.

In SemEval-2023 Task 3 (Piskorski et al., 2023a) the NAP submission (Falk et al., 2023) boosted XLM-Roberta by augmenting the training set with symmetrical back-translation and paraphrasing, achieving the best results in French and ranking in the top-3 for 7/9 languages. Alternatively, the KInIT VeraAI team (Hromadka et al., 2023) avoided augmentation: a single XLM-Roberta was fine-tuned in all languages and calibrated with dual confidence thresholds (for seen and unseen languages) and ranked first in six languages, including zero-shot surprise languages.

Domain adaptation has also emerged as a key factor. The ParlaSent (Mochtak et al., 2024) and ParlaMint datasets (Ogrodniczuk et al., 2022) provide more than a billion words of annotated parliamentary debates, enabling the development of the XLM-R-parla model[2].

Recent work also goes beyond label accuracy toward explainable persuasion detection. (Hasanain et al., 2025) release the first explanation-enhanced corpus for propaganda (Arabic and English paragraphs/tweets), where GPT-4o generated rationales are manually vetted for clarity, plausibility, faithfulness, and informativeness. The authors then finetune Llama-3.1-8B (Grattafiori et al., 2024) for label classification and generation of concise natural-language rationales.

3 Experimental setup

The primary strategy for developing the system was to fine-tune pre-trained large language models (LLMs) on datasets from previous shared tasks in related domains (Piskorski et al., 2023b, 2024). For the classification objective, we extend the pre-trained models by adding a final classification head on the top of transformer encoder. The input to the model consists of short text spans, either sentences or phrases, extracted from social networks or parliamentary debates.

For Subtask 1, the objective is binary classification: to determine whether any persuasion technique is present in the span. The model produced a single score, which was then converted into a binary label (True/False) using a threshold. For Subtask 2, which involves multi-label classification, the model's prediction head output independent scores for each persuasion technique label, allowing for the possibility that multiple techniques could co-occur in the same span.

Given this setup, we conduct additional experiments to tune the confidence thresholds, aiming to optimize overall performance. In both subtasks, we varied hyperparameters such as learning rate and batch size, and experimented with different training data configurations, including combinations of original and machine-translated examples.

The best hyperparameters used for the language models are: batch size of 64 (sBERT) and 32 (XLM-R), AdamW optimizer with 3×10^{-5} (sBERT) and 1×10^{-5} (XLM-R) learning rate and fine-tuning for 10 epochs with early stopping using cross-entropy loss.

4 Data

The input for all tasks is parliamentary debates on highly debated topics and social media posts related to the spread of disinformation in plain text format. Articles are given in five Slavic languages (Bulgarian, Polish, Croatian, Slovene and Russian). They were gathered from various sources and cover a variety of popular subjects, such as COVID-19 or the Russo-Ukrainian War, as well as abortion and migration. They were chosen primarily from the mainstream media, including some paragraphs from websites that media credibility experts have flagged as possibly disseminating false information.

We observe that some languages were associated with less reliable sources, which introduced a higher degree of bias into their datasets. For example, a large subset of the Russian-language data appear to originate from pro-government media outlets.

[2]https://huggingface.co/classla/xlm-r-parla

Model	Added languages	BG		HR		PL		RU		SI	
		Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
XLM-R	+RU, PL	**17.0**	**33.8***	**36.0***	**49.1***	**27.6***	**41.0***	12.1	17.3	**19.0***	**32.3***
sBERT	+HR$_t$	-	-	26.1	45.1	-	-	-	-	-	-
sBERT	+EN, RU, PL	14.9	33.5	27.8	46.4	19.5	31.6	8.6	14.0	14.2	27.2
sBERT	+EN, RU, PL, SI$_t$, RU$_t$	15.9	32.9	35.5	40.4	26.7	36.0	**13.3***	**18.9**	16.0	24.0

Table 1: Subtask 2 evaluation results (Macro-F1 and Micro-F1) on test set for multilingual models trained on the official dataset (BG, PL, RU, SI) with different language augmentation configurations. Added languages include both original and translated data (marked with subscript t). Asterisks for methods ranked among the top 3 in the general leaderboard.

4.1 Data Augmentation

To enrich the training data for Subtask 2, we incorporate samples from CLEF-2024 CheckThat! Task 3 (Piskorski et al., 2024). In different experimental settings, we use datasets in English (9002 samples), Russian (4138 samples) and Polish (3824 samples). These were combined with the official dataset provided by the organizers in varying proportions and configurations to assess the impact of cross-lingual augmentation on performance. No additional data was used for Subtask 1.

4.2 Data Translation

In addition to using original datasets, we also generated synthetic training data by translating "CheckThat! 2024" samples using GPT-4.1. The model is prompted to return the translations in json format. We find JSON mode of the Responses API (`"type": "json_object"`) to be useful when validating the structure of the model outputs. Since the dataset is provided as non-unique long texts and unique phrases taken from the texts, we first translate the long text and cache the translation. Then, the model is prompted to extract and return the exact translation of a given phrase. This does not always work perfectly, but the semantics of the phrase is mostly preserved. The prompt used for translation and phrase extraction can be found in Figure 1.

We select English and German as source languages due to their high resource availability and strong representation in large language models. The translated target languages are chosen based on specific weaknesses of the baseline models and gaps in the original dataset:

- Russian (RU) is selected due to relatively low validation performance, aiming to improve robustness on this language.
- Slovene (SL) is targeted because of the small size of its original training set.
- Croatian (HR) is chosen because it is entirely absent from the original training data.

In total, we obtain 1253 translated Russian samples (in addition to 4303 original ones), 1399 translated Slovene samples (compared to 82 in the original training set), 913 synthetic Croatian samples (with no original training examples). We validate $\tilde{1}00$ samples of translations into Russian and find all of them to be of a satisfactory quality. Other languages do not undergo any human evaluation.

The exact number of original, augmented, and translated data samples for each language can be found in the Table 2.

	Original	Aug	Trans	Total
BG	250	-	-	250
HR	0	-	913	913
PL	202	3824	-	4026
RU	165	4138	1253	5556
SL	82	-	1399	1481
EN	-	9002	-	9002

Table 2: Number of samples in the training set after augmentations (Aug) and machine translation (Trans).

5 Model selection

We select two families of multilingual transformer-based models for our experiments: XLM-RoBERTa-large (Conneau et al., 2019) and Slavic-BERT (Arkhipov et al., 2019).

XLM-RoBERTa-large is chosen due to its strong performance in Slavic languages such as Polish and Russian in previous multilingual shared tasks, where it consistently ranked among the top performing models (Purificato and Navigli, 2023).

Slavic-BERT is selected because it had already been pre-trained on a large Slavic-language cor-

pus, which reduces the need for additional language adaptation. Therefore, only task-specific fine-tuning is required for classification.

6 Confidence Threshold Calibration

In our experiments, we calibrate the confidence threshold separately for each language using its respective validation set. This language-specific calibration allowed us to better adapt to distributional differences across languages and enhance the overall performance of the classifiers.

For Croatian, which is not represented in the validation data, we could not calibrate the threshold in a language-specific manner. Instead, we used the average threshold derived from other languages within the same model family. This approach is motivated by the need to provide a reasonable and unbiased estimate under a zero-shot scenario.

The calibration results revealed that adjusting the confidence threshold had a significant impact on classification performance. For most languages, performance improved as the threshold decreased, up to a point beyond which it declined due to the inclusion of too many low-confidence predictions.

Optimal thresholds varied across languages and models and are provided in Table 3; however, multilingual models generally performed best with lower thresholds, particularly in zero-shot or low-resource settings. This finding aligns with previous observations that lower thresholds can compensate for reduced model confidence in scenarios with limited data coverage (Cheng et al., 2025).

While threshold optimization is beneficial to overall model performance, it can hurt the quality of the models, and this is visible in Table 1 for PL and RU, as for both languages, the model has very high Recall. This usually means that the threshold was too low (0.1 for RU) and the model predicts the positive label most of the time (99.3 recall vs. 75.5 precision for RU).

	BG	HR	PL	RU	SL
XLM-R	0.45	0.3	0.25	-	0.3
sBERT	-	-	-	0.1	-

Table 3: Optimal confidence thresholds used for label prediction in Subtask 2, selected based on validation set performance for each language and model. Thresholds are reported only for the best-performing model in each language.

7 Results

The evaluation metrics for Subtask 1 can be found in Table 4. F1 score is the main evaluation metric for this task. As it is seen from the table, the prediction quality in absolute numbers is quite high for all languages. However, in this subtask, our method based on Slavic BERT managed to score 2nd in the general leaderboard only for the Russian language with the remaining languages lagging behind. This could possibly be explained by the fact that Slavic BERT was trained on a large corpus of Russian data with a smaller percentage of other Slavic languages.

Language	Accuracy	Precision	Recall	F1-Score
BG	83.0	84.6	83.7	84.1
HR	85.1	81.1	88.2	84.5
PL	83.5	81.0	96.8	**88.2**
RU	75.3	75.5	99.3	85.8*
SL	86.4	74.1	89.2	80.9

Table 4: Subtask 1 evaluation metrics for multilingual models trained on the official dataset without data augmentation. The subtask was submitted only using SlavicBERT model. Asterisks for methods ranked among the top 3 in the general leaderboard.

Evaluation results for Subtask 2 are presented in the Table 1. We approached this task with two models, XLM-RoBERTa with official data augmentations and Slavic BERT with official and translated data augmentations. In the general leaderboard for this subtask, XLM-RoBERTa is ranked within top 3 for Croatian, Polish, and Slovene evaluated with both macro and micro F1 scores used for this task. Slavic BERT with augmentations is ranked 3rd evaluated on macro F1. It is interesting to observe that multilingual XLM-RoBERTa, which was not been specifically adapted for the Slavic languages in a way the Slavic BERT was, has scored the highest for all languages except Russian. This might indicate the advantage of model generalization across a very wide linguistic variety over a narrow domain adaptation, especially for a complicated task such as persuasion techniques classification.

8 Conclusion

The approaches we explored yielded varying results. XLM-RoBERTa performed significantly better on Subtask 2, while the solution based on SlavicBERT using augmented data achieved great scores in absolute terms for Subtask 1. This highlights how auxiliary factors can influence task perfor-

mance, such as data imbalance for certain languages or biases related to topic distribution and sentiment within the dataset.

To address such challenges, it is worth investing time in enriching the training data, which can also be achieved by translating texts from other languages. Open-sourced data from "CLEF-2024 CheckThat! Task 3" translated to minority sets increased the accuracy of the solution. Moreover, this task reminds of the importance of carefully choosing the confidence threshold for the language model's predictions.

9 Limitations

Our dataset was limited in size and exhibited a strong class imbalance, particularly for low-resource languages. There was a distinct style shift in Russian-language data, which caused a decline in performance in subtask 2. Due to computational constraints, we did not experiment with larger models or ensemble methods. Our findings are based on social media data and may not generalize well to formal or literary texts.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Bo Cheng, Jueqing Lu, Yuan Tian, Haifeng Zhao, Yi Chang, and Lan Du. 2025. Cgmatch: A different perspective of semi-supervised learning. *arXiv preprint arXiv:2503.02231*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. pages 1377–1414.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Neele Falk, Annerose Eichel, and Prisca Piccirilli. 2023. NAP at SemEval-2023 task 3: Is less really more? (back-)translation as data augmentation strategies for detecting persuasion techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1433–1446, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maram Hasanain, Md Arid Hasan, Mohamed Bayan Kmainasi, Elisa Sartori, Ali Ezzat Shahroor, Giovanni Da San Martino, and Firoj Alam. 2025. Reasoning about persuasion: Can llms enable explainable propaganda detection? *arXiv preprint arXiv:2502.16550*.

Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. KInITVeraAI at SemEval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 629–637, Toronto, Canada. Association for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. pages 16024–16036.

Maciej Ogrodniczuk, Petya Osenova, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Çağrı Çöltekin, Matyáš Kopp, and Meden Katja. 2022. ParlaMint II: The show must go on. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 1–6, Marseille, France. European Language Resources Association.

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, Alípio Jorge, Maria da Purificação Silvano, Nuno Guimarães, Ana Filipa Pacheco, and Nana Yu. 2024. Overview of the clef-2024 checkthat! lab task 3 on persuasion techniques.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. 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.

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

Antonio Purificato and Roberto Navigli. 2023. APatt at SemEval-2023 task 3: The sapienza NLP system for ensemble-based multilingual propaganda detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 382–388, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Prompt for Initial Translation:

```
Translate the following text into
{target_language}:
{document_text}

Then extract and return the exact
translations of this phrase from the
translated text:
{phrase_text}

Respond in JSON with this format:
{"translated_text":           "...",
"translated_phrases":       ["...",
"..."]}
```

Prompt for Post-translation Extraction:

```
The following text:

{document_text}

was translated into
{target_language}:

{translated_text}

Extract and return the exact
translations of this phrase from
the translated text:
{segment_text}

Respond in JSON with this format:
{"translated_phrases":       ["...",
"..."]}
```

Figure 1: Prompts for GPT-4.1 for automatic translation.