Team INSAntive at SlavicNLP-2025 Shared Task: Data Augmentation and Enhancement via Explanations for Persuasion Technique Classification

Yutong Wang and **Diana Nurbakova** and **Sylvie Calabretto**
INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205,
69621 Villeurbanne, France
{fistname.lastname}@insa-lyon.fr

Abstract

This study investigates the automatic detection and classification of persuasion techniques across five Slavic languages (Bulgarian, Croatian, Polish, Russian, and Slovenian), addressing two subtasks: binary detection of persuasion techniques in text fragments (Subtask 1) and multi-label classification of specific technique types (Subtask 2). To overcome limited training resources, we implemented a multi-level cross-lingual augmentation strategy utilizing GPT-4o for non-Slavic to Slavic conversion and intra-Slavic language migration. We employ XLM-RoBERTa architecture with two LLM-enhanced variants that use explanations to improve classification performance. The experimental results demonstrate varied performance across languages and tasks, with our approach achieving first place in the Russian subtask 1 and second place in Bulgarian subtask 2, confirming that larger parameter models excel in complex classification tasks. These findings highlight the significant potential of LLMs for enhancing multilingual classification and the persistent difficulties in ensuring consistent cross-linguistic performance.

1 Introduction and Background

This study presents our participation in the Slavic NLP 2025 shared task on the automatic detection and classification of persuasion techniques in Slavic languages. The research scope encompasses five major Slavic languages: Bulgarian (BG), Polish (PL), Croatian (HR), Slovenian (SI), and Russian (RU). The challenge comprises two closely interrelated subtasks: **Subtask 1 (Detection Task)** formulated as a binary classification problem aiming to identify the presence of one or more persuasion techniques for a given text and a list of its fragment offsets, given a predefined taxonomy of persuasion techniques. **Subtask 2 (Classification Task)** formulated as a multi-class, multi-label classification problem aiming at specify which persuasion techniques are used within a text fragment. The provided corpus contains two key categories of texts: (a) parliamentary debate transcripts on prominent social issues, and (b) social media content related to misinformation dissemination. All units of analysis are paragraph-level text fragments, enabling the research to conduct granular analysis while maintaining contextual integrity.

Starting with Da San Martino et al.'s pioneering work (Da San Martino et al., 2019) establishing an 18-category classification system, the field of automatic detection of propaganda in texts progressed through SemEval competitions (Task 11 at SemEval-2020 (Da San Martino et al., 2020), Task 3 at SemEval-2023 (Piskorski et al., 2023)) that expanded the research to multilingual contexts. Transformer-based architectures have shown significant improvements in the field (e.g. (Wu and Dredze, 2019; Arkhipov et al., 2019)). Key technological developments include XLM-RoBERTa's (Conneau et al., 2020) strong performance in cross-lingual tasks and evidence that multilingual pretrained models work effectively even for low-resource languages. Recent innovations leverage Large Language Models (LLMs) to enhance propaganda detection through several approaches: generating adversarial examples ((Hartvigsen et al., 2022)), developing explainable fake news detection (Shu et al.'s dEFEND framework (Shu et al., 2019)), and employing self-generated instructions and cloze problems for few-shot classification (Wang et al. (Wang et al., 2023); Schick and Schütze (Schick and Schütze, 2021)). This research trajectory provides theoretical foundations for using LLM-generated explanations to improve propaganda classification performance.

We propose a model that employs multi-level data augmentation to address resource scarcity in Slavic languages and utilizes an `XLM-RoBERTa`-based multi-label classification architecture, while integrating explanations generated by LLMs to en-

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 190–201
July 31, 2025 ©2025 Association for Computational Linguistics

hance both detection accuracy and interpretability.

Contribution Analysis: Prior work in cross-lingual propaganda detection has primarily focused on either data augmentation strategies (Singh et al., 2019; Lancheros et al., 2025) or explanation-enhanced models (Camburu et al., 2018) independently. Our work combines both approaches and provides detailed ablation analysis to quantify their individual contributions. Additionally, we conduct comprehensive error analysis across different language families within the Slavic group, revealing cultural and linguistic patterns that affect persuasion technique usage.

2 System overview

We present a two-phase framework for multilingual persuasion technique detection[1]. The overview of our solution is given in Fig. 1. Its *data processing* phase expands the multilingual dataset through cross-language transformation to address resource scarcity problem (see Table 1), while its model construction phase incorporates a base multi-label classifier and two architectural variants: (1) a concatenation-based integration architecture and (2) a dual-encoder cross-attention architecture, collectively forming a robust solution for propaganda detection across multiple languages.

2.1 Data Processing Stage

A dataset provided within the shared task contains quite limited resources (see Table 1). An overview of the presence of persuasion techniques across languages in the dataset is given in Appendix 5. One of the sources that could be used to expand the data is the dataset provided in SemEval 2023 Task 3 (Piskorski et al., 2023). However, in contrast to that challenge, two new persuasion techniques have been added to the taxonomy (Piskorski et al., 2025): *False Equivalence* and *Appeal to Pity*, resulting in the total of 25 techniques.

To address limited annotated data availability, we implemented a two-tier cross-lingual augmentation strategy (see Appendix 5):

1. *Non-Slavic to Slavic Conversion*: We used SemEval-2023 dataset and translated non-Slavic articles into target Slavic languages using GPT-4o, significantly expanding Russian and Polish training samples while preserving persuasion technique structures.
2. *Intra-Slavic Migration*: For Croatian, Slovenian, and Bulgarian—languages entirely absent from the SemEval-2023 dataset—we translated existing Russian and Polish articles using GPT-4o, minimizing semantic shifts and rhetorical structure deformations.

For the newly added persuasion techniques, we employed a guided generation method, creating 50 original articles per target language using GPT-4o while ensuring consistency with the original dataset's style. All prompts used are provided in Appendix 5.

We implemented systematic data processing strategies including text normalization, length filtering (excluding sequences >1000 characters), tokenization using `XLM-RoBERTa` dedicated tokenizer, and multi-label encoding to convert label strings into multi-hot vector representations.

2.2 Model Building Stage

Our framework leverages Transformer architecture with targeted optimization strategies to identify complex persuasion patterns. Based on multilingual processing requirements, we employ `XLM-RoBERTa` (Conneau et al., 2020) as the core model, exploring two variants: `XLM-RoBERTa-base` (a standard variant with approximately 125M parameters); and `XLM-RoBERTa-large` (an expanded variant with approximately 355M parameters, used to enhance model capacity and performance ceiling).

The architecture primarily consists of the following components: a pre-trained `XLM-RoBERTa` encoder for extracting deep contextual text representations; dedicated classification layers generating logits values for each persuasion technique; and Sigmoid activation functions converting logits values into independent probability of existence for each category.

The core prediction formula of the model can be expressed as: $P(y_i = 1|x) = \sigma(f_i(x))$, where σ represents the Sigmoid activation function, $f_i(x)$ is the output logit for class i, x is the input text, and $P(y_i = 1|x)$ represents the probability of category i existing in input x.

To address the inherent class imbalance problem (see Appendix 5), we adopt binary cross-entropy (BCE) loss with adaptive class weights. BCE loss is defined as: $\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$, where y_i denotes the true label, $\hat{y}_i$ represents the predicted label, and N is the total number of

[1] The code will be available on https://github.com/dalanzuipang/INSAntive_at_SlavicNLP-2025

			BG	HR	PL	RU	SL	EN
TRAIN	BEFORE	# articles	20	-	15	27	15	-
		# spans	363	-	289	239	108	-
		# spans with persuasion	168 (46.3%)	-	195 (67.5%)	166 (69.5%)	58 (53.7%)	-
	AFTER	# articles	368	368	742	844	368	526
TEST		# articles	59	10	38	63	36	-
		# spans	1361	74	729	590	487	-

Table 1: Dataset statistics before and after data augmentation

categories. To balance the contribution of each category, we use a weight adjustment mechanism inversely proportional to category frequency:

$$w_i = \min\left(\frac{n_{neg,i}}{n_{pos,i}} \cdot \text{ratio}, w_{max}\right) \tag{1}$$

where $n_{pos,i}$ and $n_{neg,i}$ represent the number of positive and negative samples for category i respectively, ratio is an adjustment proportion factor (set to 3.0), and w_{max} is the maximum weight limit (set to 30.0) to avoid numerical instability.

2.3 Model Enhancement via Explanations

To enhance detection performance, we enhance the model with LLM-generated explanations (see Appendix 5 for prompts). To do so, we developed two innovative variants of architecture: (1) concatenation-based ensemble and (2) dual-encoder cross-attention architecture.

Concatenation-based ensemble architecture: First, we obtain descriptive explanations for each text fragment in training data. Then, using a dedicated separator token `[SEP]` we concatenate the original text with its corresponding explanation. We truncate explanation content to a maximum of 128 tokens to manage input length while preserving core information. The model architecture for this variant remains consistent with the basic multi-label classifier, with only the input processing pipeline modified to accommodate the combined text-explanation format.

Dual-Encoder Cross-Attention Architecture This method processes text and explanations through independent encoders and integrates their representations using a refined cross-attention mechanism that consists of three main steps:

1. **Parallel encoding process:** The text encoder specifically processes original text content, while the explanation encoder specifically processes LLM-generated explanation content.
2. **Cross-attention integration:** Implementing bidirectional information flow between text and explanation representations, including text-to-explanation attention and explanation-to-text attention.
3. **Multi-dimensional feature fusion:** The model synthesizes four complementary feature representations: original text representation, original explanation representation, text-attentive-to-explanation representation, and explanation-attentive-to-text representation.

3 Experimental Setup

Our model implementation leverages PyTorch and PyTorch Lightning frameworks for structured and efficient training. To enhance training stability and performance, we employed several optimization techniques including gradient checkpointing to reduce memory requirements, gradient clipping to prevent explosion phenomena, gradient accumulation to achieve large-batch training while circumventing memory limitations, linear learning rate scheduling with warm-up for stabilizing initial training, and epsilon stabilization to prevent numerical instability in loss calculations. We used two model variants with optimized hyperparameters: (1) `XLM-RoBERTa-base`: Batch size: 8; Gradient accumulation steps: 4; Learning rate: 1×10^{-5}; Warm-up steps: 1000; (2) `XLM-RoBERTa-large`: Batch size: 4; Gradient accumulation steps: 8; Learning rate: 5×10^{-6}; Warm-up steps: 2000.

For explanation-enhanced methods, we applied multiple optimizations: (a) Input length management: maximum sequence length for text set to 256 tokens, explanations limited to 128 tokens, achieving a balance between computational resources and model expressive capacity; (b) Elastic inference: support for selectively providing explanations during inference, automatically reverting to using only original text input when no explanation is available; (c) Enhancement rather than dependence: ensuring the model architecture benefits from explanations without over-reliance, maintaining robust performance even when explanations are unavailable.

4 Results and Analysis

We compare five model configurations, focusing particularly on the integration of explanations with our XLM-RoBERTa architecture: (1) fine-tuned XLM-RoBERTa-base, (2) XLM-RoBERTa-base+Concatenation, (3) XLM-RoBERTa-base+Dual encoder, (4) fine-tuned XLM-RoBERTa-large (5) XLM-RoBERTa-large+Concatenation. Here, we report only the best configuration results.

Our system demonstrated varied performance across the evaluation metrics, with distinct strengths in specific language-subtask combinations (see Tables 2 and 3).

We observed consistent cross-linguistic patterns in performance metrics. In subtask 1, the system demonstrated high precision (0.8454-0.9355) but varied recall (0.5223-0.8784), indicating a conservative classification approach favouring high-confidence identifications while potentially overlooking positive instances. In subtask 2, micro F1 scores (0.1969-0.4081) consistently exceeded macro F1 scores (0.1365-0.2620), revealing better performance on frequent persuasion techniques compared to rare categories. The 65.2% performance gap between the highest (Polish: 0.2671) and lowest (Slovenian: 0.1388) performing languages reflects the inherent diversity within the Slavic language family and varying resource availability.

4.1 Per-Class Performance Analysis and Cross-Linguistic Patterns

To provide deeper insights into our system's behaviour across different persuasion techniques and languages, we conducted comprehensive per-class evaluations for all 25 persuasion technique categories across the five Slavic languages and five model configurations (See Appendix 5). Our detailed evaluation across all configurations reveals clear architectural preferences for different languages.

Configuration effectiveness ranking:

1. **Configuration 3 (Dual-Encoder)**: Average F1 = 0.1922, optimal for 4/5 languages
2. **Configuration 2 (Concatenation)**: Average F1 = 0.1769, optimal for 1/5 languages
3. **Configuration 5 (Large+Concat)**: Average F1 = 0.1474
4. **Configuration 1 (Base)**: Average F1 = 0.1431
5. **Configuration 4 (Large)**: Average F1 = 0.1394

Notably, larger parameter models (Configurations 4 and 5) show a consistent pattern of high precision but low recall, suggesting they adopt more conservative prediction strategies. This precision-recall trade-off indicates that while larger models make fewer false positive predictions, they miss a significant number of true persuasion techniques.

Error analysis revealed several systematic failure modes:

Configuration-Specific Errors:

- Large models (Config 4, 5) consistently under-predict rare techniques (Appeal to Pity, False Equivalence)
- Base model (Config 1) shows poor performance on nuanced techniques requiring contextual understanding
- Dual-Encoder (Config 3) occasionally over-relies on explanation content, leading to false positives when explanations are imperfect

Language-Specific Challenges:

- **Slovenian**: Severe data sparsity leads to poor generalization for infrequent techniques
- **Russian**: Morphological complexity creates false pattern matches
- **Croatian**: Limited label coverage (missing 2 techniques) affects overall system robustness

The experimental results reveal important insights. Language performance variations highlight the need for language-specific model adjustments, especially for resource-limited languages like Slovenian. Low performance in terms of F1-scores in subtask 2 shows that the latter remains an open challenge. The system's better performance in terms of ranking on classification tasks compared to detection tasks demonstrates its ability to distinguish between persuasion techniques, though binary decision making need refinement. The significant gap between micro and macro metrics in multi-label classification emphasizes the need to address class imbalance issues. Overall, these findings showcase both the potential of LLMs for enhancing multilingual classification and the ongoing challenges in achieving consistent performance across diverse languages and technique categories.

5 Conclusion

This paper introduces a framework for persuasion technique detection across five Slavic languages that combines cross-lingual data augmentation, XLM-RoBERTa architecture, and explanation in-

Language	Rank	Config	Accuracy	Precision	Recall	F1-score
Russian	**1/7**	3	0.8051	0.8647	0.8784	0.8715
Croatian	4/6	2	0.9054	0.9355	0.8529	0.8923
Polish	6/7	3	0.8436	0.8799	0.8723	0.8761
Bulgarian	7/7	2	0.8097	0.8802	0.7497	0.8097
Slovenian	6/7	2	0.8152	0.8454	0.5223	0.6457

Table 2: Ranking and performance metrics for Subtask 1 (Binary Detection)

Language	Micro-Rank	Macro-Rank	Config	Accuracy	Micro F1	Macro F1
Russian	**1/6**	2/6	4	0.1932	0.2958	0.1779
Bulgarian	2/7	2/7	3	0.3865	0.3440	0.2082
Polish	3/7	4/7	2	0.3251	0.4081	0.2620
Slovenian	3/7	4/7	3	0.4949	0.1969	0.1365
Croatian	5/7	6/7	2	0.5270	0.2933	0.1778

Table 3: Rankings and performance metrics for Subtask 2 (Multi-label Classification)

tegration mechanisms. The approach achieved top rankings in Russian and Bulgarian subtasks. Key findings demonstrate that: (1) larger models more effectively capture persuasive language patterns, (2) integrating LLM-generated explanations via cross-attention mechanisms significantly improves performance, and (3) cross-lingual augmentation effectively addresses data scarcity in low-resource languages within the same language family. Future work will explore knowledge base integration, advanced cross-lingual transfer techniques, and specialized architectures for logical relationship modelling in persuasive text.

Acknowledgments

Yutong Wang is supported by the China Scholarship Council scholarship for Ph.D. program at INSA Lyon, France. File No. 202308120039.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31, pages 9539–9549. Curran Associates, Inc.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 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)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

B.S. Lancheros, G. Corpas Pastor, and R. Mitkov. 2025. Data augmentation and transfer learning for cross-lingual named entity recognition in the biomedical domain. *Language Resources and Evaluation*, 59:665–684.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. 2025. Persuasion Technique Taxonomy used in the Shared Task on the Dectection and Classification of Persuasion Techniques in Texts for Slavic Languages.

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, Toronto, Canada. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *Preprint*, arXiv:1905.11471.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Appendix A: Persuasion Technique Detection Framework

Appendix B: Occurrence of Persuasion Techniques across Languages

In this Section, we present the occurrence of persuasion techniques among all languages (Fig. **??**) and for each individual language (Fig. 3 - Fig. 6) based on the TRAIN set of the Shared Task. Figure 7 provides a heatmap of technique frequencies across languages.

Though the number of articles is relatively small, we can still do some observations. Most notably, persuasion techniques demonstrate skewed distribution. A small number of techniques (*Loaded Language*, *Questioning the Reputation*, *Doubt*) account for a disproportionately large share of the total occurrences. Thus *Loaded Language* (171 total instances) is the most prevalent technique overall, especially dominant in Polish and Slovene. This indicates widespread reliance on emotionally charged language to persuade. *Questioning the Reputation* (138 instances) also has a strong presence in all languages, with Bulgarian and Polish contributing most heavily, suggesting these cultures frequently use credibility attacks. *Doubt* (136 instances) is strongly present in Bulgarian (12.2%) and Slovenian (8.9%), showing the importance of creating uncertainty about opposing viewpoints.

However, each language shows different patterns of technique usage. While *Loaded Language* has the highest raw count, its proportional use varies significantly, suggesting different cultural norms around emotional language. Bulgarian persuasion relies heavily on direct confrontation techniques: reputation questioning, name-calling, doubt. Russian persuasion emphasises emotional appeals (fear, values) and oversimplification techniques. Polish shows the most balanced approach though maintaining a skewed distribution, suggesting more varied persuasion strategies. Slovenian persuasion focuses on authority and doubt by prominently using *Appeal to Authority*, *Appeal to Values*, *Doubt* and *Loaded Language*. We can also observe few notable contrasts: (a) *Appeal to Values* is barely used in Bulgarian (1.4%) but heavily employed in Polish (10.2%), Russian (9.3%), and Slovenian (9.2%); (b) *Name Calling-Labeling* is much more prevalent in Bulgarian (11.6%) than in other languages; (c) *Conversation_Killers* are completely absent in Slovenian (0.0%) but used in other languages, particularly Polish (6.4%).

The distributions of persuasion techniques after data augmentation are given in Figures 8-11. Note that augmented dataset contains data for Croatian and English. *Loaded Language* remains dominant across all languages with the highest proportion in English. We note that this augmented data suggests greater similarity between Bulgarian, Slovenian, Croatian, and Russian than the original Shared Task dataset. For instance, the differences like Bulgarian's strong reliance on confrontational techniques are less pronounced in the augmented dataset. Due to the use of translation, we note some shifts in distributions, such as: *Appeal to Values*

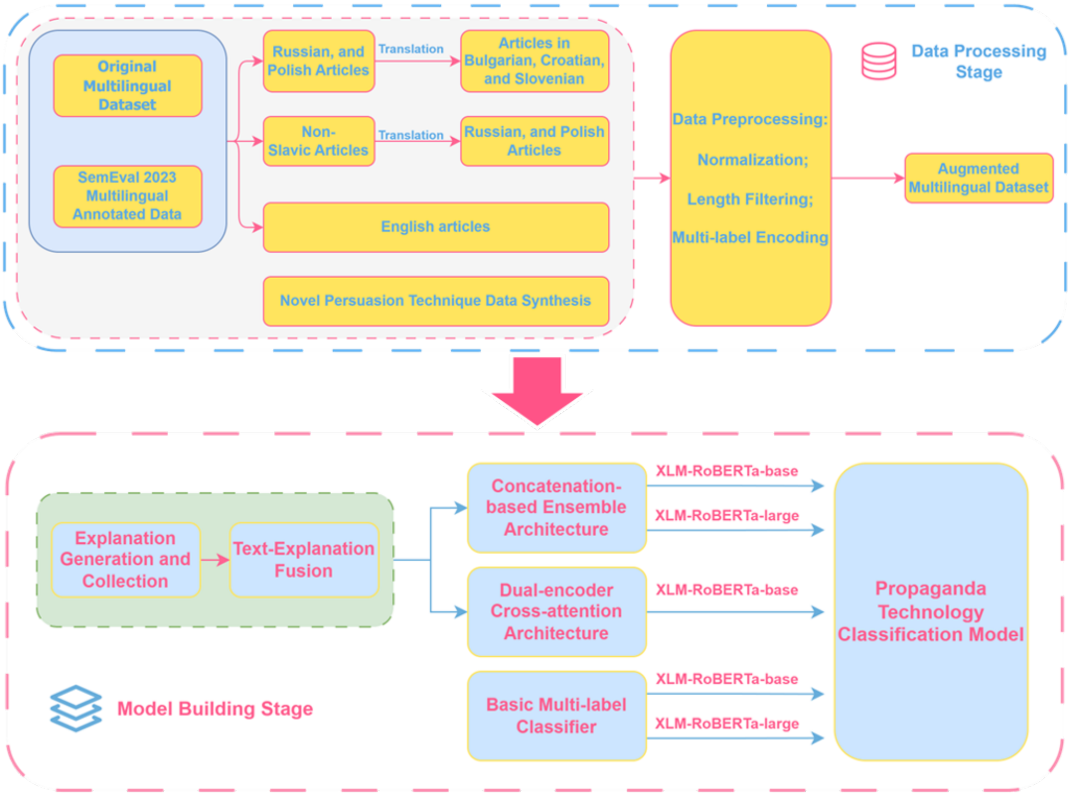

Original Multilingual Dataset
SemEval 2023 Multilingual Annotated Data
Russian, and Polish Articles
Translation
Articles in Bulgarian, Croatian, and Slovenian
Non-Slavic Articles
Translation
Russian, and Polish Articles
English articles
Novel Persuasion Technique Data Synthesis
Data Preprocessing:
Normalization;
Length Filtering;
Multi-label Encoding
Data Processing Stage
Augmented Multilingual Dataset
Explanation Generation and Collection
Text-Explanation Fusion
Concatenation-based Ensemble Architecture
Dual-encoder Cross-attention Architecture
Basic Multi-label Classifier
XLM-RoBERTa-base
XLM-RoBERTa-large
XLM-RoBERTa-base
XLM-RoBERTa-base
XLM-RoBERTa-large
Propaganda Technology Classification Model
Model Building Stage

Figure 1: Overview of our persuasion technique classification model

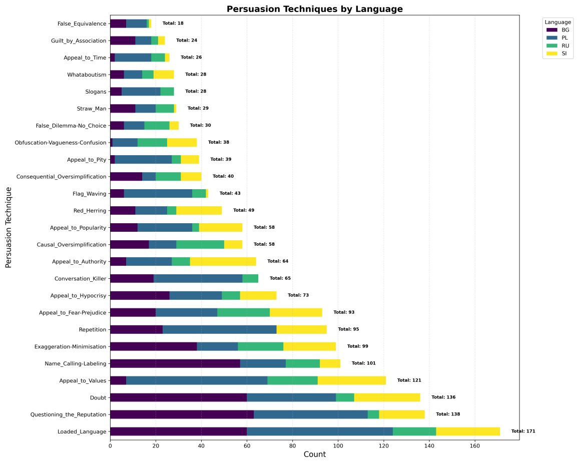


Figure 2: Occurrence of persuasion techniques in the TRAIN set

appears less significant in the augmented dataset, *Conversation Killer* shows more presence in Slovenian, Croatian, and Bulgarian than in the original dataset, *Appeal to Authority* ranks lower across all languages. Across all languages, a small set of techniques (*Loaded Language*, *Name Calling*, *Doubt*, *Questioning Reputation*) forms the core persuasion toolkit, accounting for roughly 50-60% of all persuasive techniques. Interestingly, the introduction of English provides a curious contrast point, revealing potential Western vs. Slavic differences in persuasion strategies, in particular: stronger reliance on patriotic appeals (*Flag Waving*), lower emphasis on creating doubt, and higher usage of emotional language. Another interesting observation is that Bulgarian, Croatian, Slovenian seem to show similar patterns suggesting a potential cultural cluster of South Slavic rhetorical approaches.

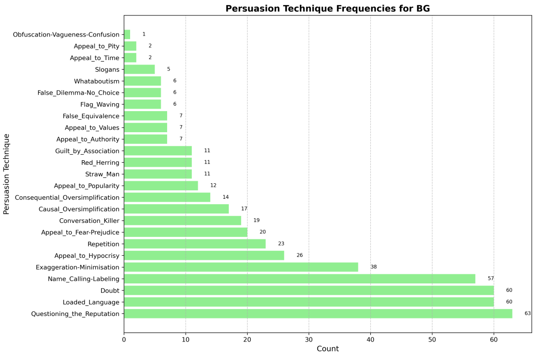


Figure 3: Bulgarian (TRAIN)

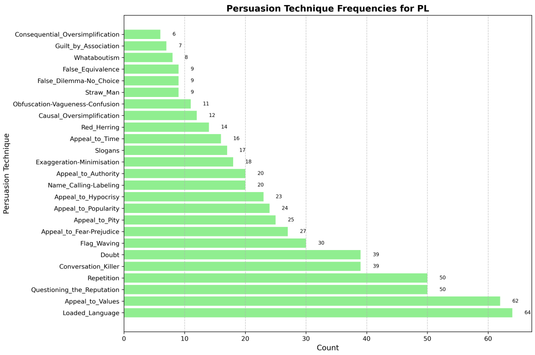


Figure 4: Polish (TRAIN)

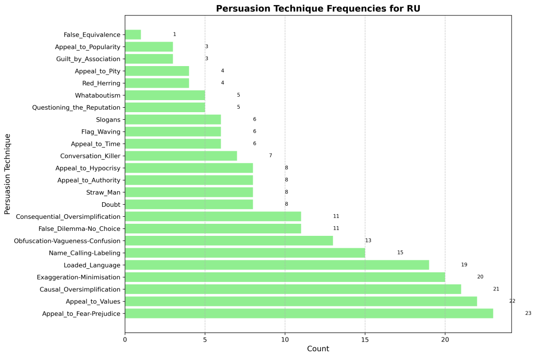


Figure 5: Russian (TRAIN)

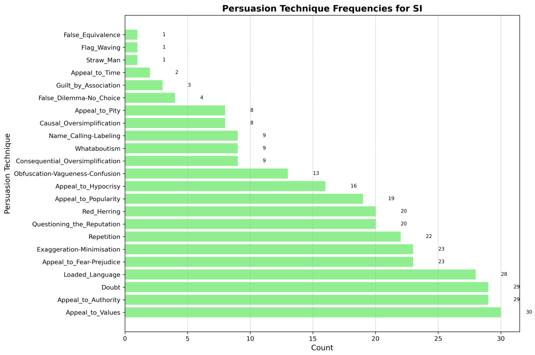


Figure 6: Slovene (TRAIN)

Appendix C: Prompts

In this Section, we provide prompts used in our approach.

Document Translation Prompt

```
You are a professional translator.
Translate the following text from {
source_lang} to {target_lang}. Maintain
the original format and paragraph
structure. Translate everything
accurately and completely.
```

News Article Generation with Persuasion Techniques Prompt

```
System Prompt:
You are a professional multilingual
content writer, skilled at creating
various types of articles according to
requirements.

User Prompt:
Please write a news article of 800-1000
words based on the topic "{topic}". In
the article, please include at least 1
paragraphs that use the propaganda
technique "{label_info['label_name']}"
({label_info['label_english']}).
The definition of {label_info['
label_english']} is: {label_info['
definition']}
Requirements:
    1. The article should have a title,
introduction, body, and conclusion
    2. Clearly mark paragraphs that use
the "{label_info['label_english']}"
technique by adding comments before and
after the paragraph <!-- {label_info['
label_english']} -->
    3. Please ensure the article overall
 looks like a real discussion of issues
or opinion piece
    4. The rest of the article should
use reasonable arguments and logic
    5. The article must be written in {
language_name}

Appeal to Pity:
Appeal to Pity: A technique that evokes
feelings of pity, sympathy, compassion
or guilt in audience to distract it from
 focusing on evidence, rational analysis
 and logical reasoning, so that it
```

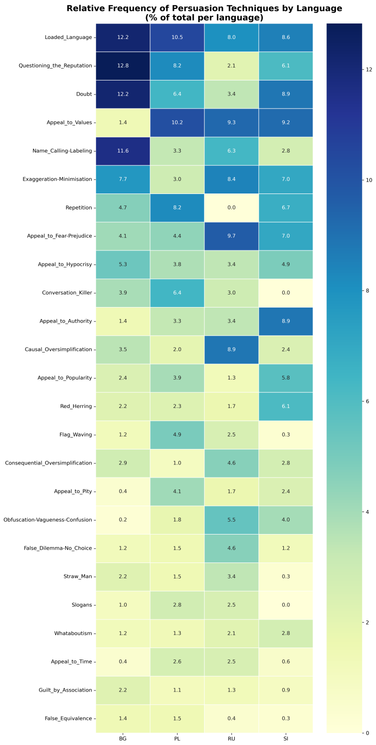
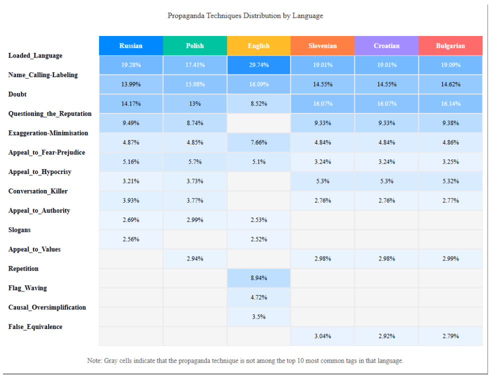


Figure 7: Heatmap of persuasion techniques across Slavic languages in the TRAIN set



Propaganda Techniques Distribution by Language

	Russian	Polish	English	Slovenian	Croatian	Bulgarian
Loaded_Language	19.28%	17.41%	29.74%	19.01%	19.01%	19.09%
Name_Calling-Labeling	13.99%	15.98%	16.09%	14.55%	14.55%	14.62%
Doubt	14.17%	13%	8.52%	16.07%	16.07%	16.14%
Questioning_the_Reputation	9.49%	8.74%		9.33%	9.33%	9.38%
Exaggeration-Minimisation	4.87%	4.85%	7.66%	4.84%	4.84%	4.86%
Appeal_to_Fear-Prejudice	5.16%	5.7%	5.1%	3.24%	3.24%	3.25%
Appeal_to_Hypocrisy	3.21%	3.73%		5.3%	5.3%	5.32%
Conversation_Killer	3.93%	3.77%		2.76%	2.76%	2.77%
Appeal_to_Authority	2.69%	2.99%	2.53%			
Slogans	2.56%		2.52%			
Appeal_to_Values		2.94%		2.98%	2.98%	2.99%
Repetition			8.94%			
Flag_Waving			4.72%			
Causal_Oversimplification			3.5%			
False_Equivalence				3.04%	2.92%	2.79%

Note: Gray cells indicate that the propaganda technique is not among the top 10 most common tags in that language.

Figure 8: Distribution of persuasion techniques across languages after data augmentation

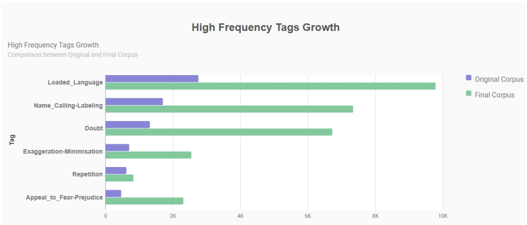


Figure 9: High frequency techniques growth

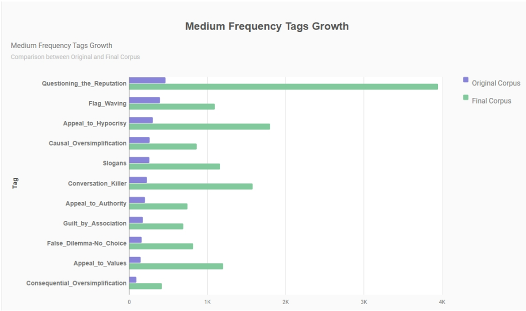


Figure 10: Medium frequency techniques growth

accepts the speaker's conclusion as truthful solely based on soliciting the aforementioned emotions. It is an attempt to sway opinions and fully substitute logical evidence in an argument with a claim intended to elicit pity or guilt.

False Equivalence:

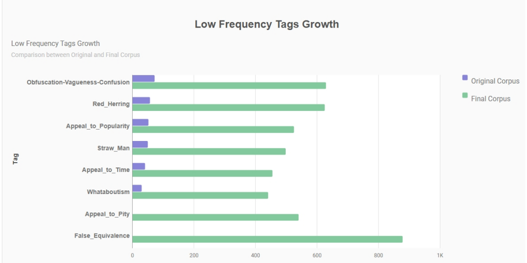


Figure 11: Low frequency techniques growth

False Equivalence: A technique that attempts to treat scenarios that are significantly different as if they had equal merit or significance. In particular, an emphasis is being made on one specific shared characteristic between the items of comparison in the argument that is way off in the order of magnitude, oversimplified, or just that important additional factors have been ignored. The introduction of the certain shared characteristics of the scenarios is then used to consider them equal. This technique has the following logical form: A and B share some characteristic X. Therefore, A and B are equal or equivalent in value, merit or significance.

Justification Prompts

Target Label Justification Prompt

Please analyze the propaganda technique in the following text, specifically focusing on "{target_label}":
Text: {row['text']}

The true label includes "{target_label}", but the predicted label does not. Please analyze the following questions in English, and combine your answers into a coherent paragraph with a maximum length of 150 words:

1. What are the specific reasons why this text contains {target_label}?
2. What key words or phrases in the text support this judgment?
3. How are the typical characteristics of {target_label} reflected in the text?
4. What additional features should the model pay attention to in the text to more accurately identify this technique?

Note: All answers must be combined into a single paragraph without bullet points or numbering, ensuring the content is coherent and does not exceed 150 words.

Wrongly Predicted Label Justification Prompt

Please analyze the error in predicting the propaganda technique "{target_label}" in the following text:
Text: {row['text']}

The predicted label includes "{target_label}", but the true label does not. Please analyze the following questions in English, and combine your answers into a coherent paragraph with a maximum length of 150 words:

1. What are the specific reasons why this text does not contain {target_label}?
2. What key words or phrases in the text support this judgment?
3. How are the typical characteristics of {target_label} reflected in the text?
4. What misconceptions or error patterns might the model have when identifying {target_label}?

Note: All answers must be combined into a single paragraph without bullet points or numbering, ensuring the content is coherent and does not exceed 150 words.

Correctly Predicted Label Justification Prompt

Please analyze the correctly identified propaganda technique "{target_label}" in the following text:
Text: {row['text']}

Both the true label and predicted label include "{target_label}". Please analyze the following questions in English, and combine your answers into a coherent paragraph with a maximum length of 150 words:

1. What are the specific reasons why this text contains {target_label}?

```
    2. What key words or phrases in the
text support this judgment?
    3. How are the typical
characteristics of {target_label}
reflected in the text?
    4. What was the key to the model
correctly identifying this technique?

Note: All answers must be combined into
a single paragraph without bullet points
 or numbering, ensuring the content is
coherent and does not exceed 150 words.
```

Confusion Justification Prompt

```
Please analyze the confusion between
propaganda technique labels in the
following text:
Text: {row['text']}

The true label is "{target_label}", but
it was predicted as "{confused_label}".
Please analyze the following questions
in English, and combine your answers
into a coherent paragraph with a maximum
 length of 150 words:
    1. Why does this text better fit {
target_label} rather than {
confused_label}?
    2. What are the key differences
between these two techniques?
    3. What might be the reasons for the
 model confusing these two techniques?

Note: All answers must be combined into
a single paragraph without bullet points
 or numbering, ensuring the content is
coherent and does not exceed 150 words.
```

Appendix D: Detailed Performance

In Table 4, we lists the overall performance breakdown for each language-configuration combination.

Language	**Config**	**F1 Score**	**Precision**	**Recall**	**Architecture Type**
Polish (PL)	PL_3*	**0.2671**	0.2495	**0.3284**	Dual-Encoder
	PL_2	0.2535	0.3211	0.2514	Concatenation
	PL_5	0.2239	**0.3866**	0.1910	Large+Concat
	PL_1	0.2218	0.3014	0.2054	Base Model
	PL_4	0.2113	0.3589	0.1805	Large Model
Bulgarian (BG)	BG_3*	**0.2132**	0.2397	**0.2550**	Dual-Encoder
	BG_5	0.1952	0.3334	0.1828	Large+Concat
	BG_2	0.1938	0.3137	0.1994	Concatenation
	BG_4	0.1836	**0.3623**	0.1684	Large Model
	BG_1	0.1537	0.2612	0.1440	Base Model
Croatian (HR)	HR_2*	**0.1824**	**0.2835**	0.1765	Concatenation
	HR_3	0.1601	0.1731	0.1659	Dual-Encoder
	HR_1	0.0978	0.1964	0.0828	Base Model
	HR_5	0.0933	0.1569	0.0905	Large+Concat
	HR_4	0.0851	0.1497	0.0853	Large Model
Russian (RU)	RU_3*	**0.1817**	0.1448	**0.3307**	Dual-Encoder
	RU_2	0.1657	0.1809	0.2337	Concatenation
	RU_4	0.1639	0.1683	0.2241	Large Model
	RU_5	0.1548	0.1573	0.2269	Large+Concat
	RU_1	0.1465	0.1462	0.1998	Base Model
Slovenian (SI)	SI_3*	**0.1388**	0.1913	0.2100	Dual-Encoder
	SI_1	0.0956	**0.2302**	0.0933	Base Model
	SI_2	0.0889	0.1934	0.1122	Concatenation
	SI_5	0.0698	0.1397	0.0763	Large+Concat
	SI_4	0.0532	0.0980	0.0625	Large Model

Table 4: Detailed per-class performance analysis by language and configuration (*Best configuration for each language)