Fine-Tuned Transformer-Based Weighted Soft Voting Ensemble for Persuasion Technique Classification in Slavic Languages

Mahshar Yahan, Sakib Sarker, Mohammad Amanul Islam
Department of Computer Science and Engineering
Uttara University, Bangladesh
mahshar@uttara.ac.bd, {sakib.sarker, amanul.islam}@uttarauniversity.edu.bd

Abstract

This paper explores detecting persuasion techniques in Slavic languages using both single transformer models and weighted soft voting ensemble methods. We focused on identifying the presence of persuasion in Bulgarian, Polish, Slovene, and Russian text fragments. We have applied various preprocessing steps to improve model performance. Our experiments show that weighted soft voting ensembles consistently outperform single models in most languages, achieving F1-scores of 0.867 for Bulgarian, 0.902 for Polish, and 0.804 for Russian. For Slovene, the single SlovakBERT model performed best with an F1-score of 0.823, just ahead of the ensemble. These results demonstrate that combining monolingual and multilingual transformer models is effective for robust persuasion detection in low-resource Slavic languages.

1 Introduction

Persuasion techniques are widely used in today's digital world, especially in political debates and on social media. These techniques aim to influence people's opinions and decisions, sometimes in ways that are not always fair or truthful. Because of their impact, it is important to develop methods for automatically detecting and analyzing these techniques in different languages and contexts.

The Slavic NLP 2025 Shared Task (Piskorski et al., 2025) focuses on this challenge by inviting participants to build systems that can identify persuasion techniques in five Slavic languages: Bulgarian, Polish, Croatian, Slovene, and Russian. The task uses real-world data from two main sources: parliamentary debates on controversial topics and social media posts that often spread disinformation. Both sources are known to contain a wide variety of persuasive strategies.

A study (Bassi et al., 2024) has shown that persuasive content online can directly impact democratic processes by shaping public opinion and even influencing election outcomes. Another study (Traberg et al., 2024) experimented with 20,477 participants and found that social cues like endorsements or high numbers of likes significantly increased belief in misinformation [M[1] = 2.83 vs. M = 2.23, p[2] < 0.001].

In this paper, we focus on the first subtask of the shared task, which is a binary classification problem. The goal is to decide whether any persuasion technique is present in a given fragment of text.The organizers have provided annotated datasets for the trial, training, and test phases. To tackle this problem, we employ transformer-based models, which have demonstrated strong performance on similar natural language processing tasks. Since this is a downstream classification task, we focus on encoder-based architectures such as BERT, RoBERTa, and XLMNet. Model performance is primarily evaluated using the F1 score.

The major contributions of our research work are as follows-

- We proposed both single-model and weighted ensemble approaches using transformer models for Slavic languages to achieve effective results.
- We conducted a series of experiments on the dataset and provided a thorough analysis of their performance.

The experimentation details have been provided in the GitHub repository. [3]

2 Related Work

The detection and classification of persuasion techniques in text has gained increasing attention

[1] M = Average belief in misinformation
[2] p = Probability of observed result happened by chance
[3] https://github.com/mahshar-yahan/SalvicNLP-2025/tree/main/ST_1

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 217–223

in recent years because of the amount of misinformation and manipulative content online.

A large-scale dataset and baseline systems(Martino et al., 2020) were introduced in 2020 to detect propaganda in news articles containing 18 persuasion techniques, addressing both the detection and classification of technique spans. Their work laid the foundation for later shared tasks and research on identifying persuasion techniques. Recent studies(Dimitrov et al., 2021) have worked on creating detailed systems that identify different rhetorical and manipulative strategies in both paragraphs and smaller parts of text.

A recent study(Nikolaidis et al., 2023) showed that both monolingual and multilingual BERT(Bidirectional Encoder Representations from Transformers) models work well, especially for languages like Polish and Russian. Authors of the paper(Scannell et al., 2021) on COVID-19 show remarkable result detecting persuasion about vaccination. This study applied RoBERTa(Robustly Optimized BERT Approach), Grover, and ELMo(Embeddings from Language Models) to detect persuasion from social media and news articles.

Another study (Nikolaidis et al., 2023) detects persuasion techniques in Polish and Russian news using transformer models: monolingual HerBERT (Polish), RuRoBERTa (Russian), and multilingual XLM-RoBERTa. The main objective of this paper is to find which multilingual model is most effective to detect persuasion on social media. A recent research detects propaganda techniques in memes (including Bulgarian) using the HPT[4] hierarchical text classification model (Ghahroodi and Asgari, 2024), achieving top results for English text-only sub-tasks and competitive performance for Bulgarian.

In summary, new transformer models and improved multilingual datasets are advancing to detect persuasion and manipulation in different languages and online platforms. These developments are important for fighting misinformation and understanding how people are influenced online (Lazer et al., 2018).

3 Dataset

The dataset analyzed in this study originates from the Slavic NLP 2025 Shared Task 1(Piskorski et al., 2025), which focuses on the detection of persuasion techniques in text across multiple Slavic languages. The provided dataset contains four languages, which are Bulgarian, Polish, Slovene, and Russian. The texts in the dataset mainly contain parliamentary debates on highly discussed topics and social media posts related to the spread of disinformation. In addition to the main training and testing sets, a small trial set of raw text was also provided during the trial period. We merged the train and trial sets to create the total dataset. From the total dataset, we use 90% for training and 10% for evaluation. The dataset is divided into trial, training, and testing sets, as shown in the table 1.

Split	Bulgarian	Polish	Slovene	Russian
Trail	75	27	9	18
Train	363	289	108	239
Test	438	729	487	590

Table 1: Language-wise distribution in the dataset

4 Methodology

In this section, we give a clear overview of the methods used to analyze the dataset. First, we preprocessed and tokenized the data to prepare it for modeling. Next, we trained individual models using both the trial and training sets. After evaluating their performance, we selected the best models upon their F1 score and combined their predictions using a weighted ensemble approach. The following diagram 1 illustrates the process for both training single models and using an ensemble of multiple models to detect persuasion in the text.

4.1 Data Preprocessing

Several preprocessing steps have been applied on the given dataset of different language to achieve optimal outcomes. After studying several papers, we have decided that the preprocessing steps should include removing punctuation and standardizing text formatting. Each step aims to improve the model's capacity to process linguistic patterns more effectively.

4.1.1 Punctuation Removal

Through removing punctuation, the model finds text patterns more easily. As some of the data in this task was collected from social media, where punctuation is often used in non-standard or inconsistent ways. By removing such symbols, the

[4]Hierarchy-aware Prompt Tuning

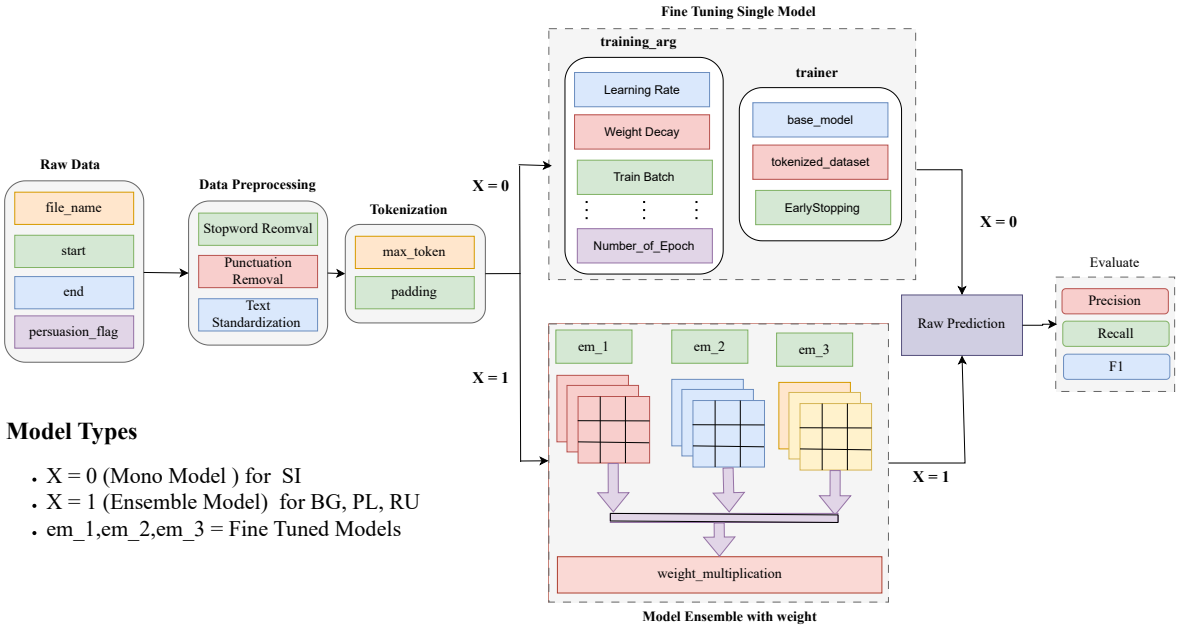

Figure 1: Methodological Workflow for Detecting Persuasion in Slavic Language Using both Single Model and Ensemble models

text becomes cleaner and less noisy, which makes the input more uniform to model(Scannell et al., 2021).

Before Removal: ...pomagając swoim bliskim, kiedy są w kryzysie?

After Removal: pomagając swoim bliskim kiedy są w kryzysie

In this example, the punctuation present before removal does not significantly affect the sentiment, so we opted to remove it.

4.1.2 Text Standardization

Since most of the data comes from parliamentary debates, it includes human-like speech with frequent line breaks. By removing these speech-like structures and line breaks, we convert the text into plain and continuous text that is suitable for model training.

Before Standardization:

Poseł Dorota Olko:
Panie Marszałku! Wysoka Izbo!
Posłanka Maria Żukowska mówiła już o tym,
o czym jest ta ustawa. Ja zacznę od dwóch historii.

After Standardization:

Poseł Dorota Olko Panie Marszałku Wysoka Izbo Posłanka Maria Żukowska mówiła już o tym o czym jest ta ustawa Ja zacznę od dwóch historii

4.2 Tokenization

As we are using BERT-based (Devlin et al., 2019) models,so BERT tokenizers are used for splitting the text into tokens. We set the maximum sequence length to 512. If a sentence is shorter than 512 tokens, we add padding to the right to reach this length. We do not use dynamic padding because most of our sentences are already close to the maximum length.

4.3 Single Model Train

In our single model (X=0) training, we implemented several strategies to optimize performance and efficiency. Early stopping was utilized to prevent unnecessary training epochs and reduce training time by halting the process when no further improvements were observed. We incorporated weight decay as a regularization technique to minimize the risk of overfitting. Additionally, the model was trained using batched data with a batch size of 2, which helps stabilize gradient updates and efficiently utilizes computational resources. The reason behind selecting a smaller batch size is the size of the dataset. Using a small batch size helps ensure that the model sees more parameter updates per epoch, which can be beneficial for learning from scarce data.

Language	Model	Weight	F1
Bulgarian	BERTić* (Ljubešić and Lauc, 2021)	NA	0.861
	baseline	NA	0.88
	Proposed Ensemble	[0.6, 0.2, 0.2]	**0.867**
Polish	Polish-roberta(Semary et al., 2023)	NA	0.897
	baseline	NA	0.90
	Proposed Ensemble	[0.4, 0.1, 0.5]	**0.902**
Slovene	**SlovakBERT** (Pikuliak et al., 2021)	NA	**0.823**
	baseline	NA	0.85
	Proposed Ensemble	[0.4, 0.1, 0.5]	0.815
Russian	Conversational RuBERT (Galimzianov and Vyshegorodtsev, 2024)	NA	0.778
	baseline	NA	0.83
	Proposed Ensemble	[0.5, 0.25, 0.25]	**0.804**

Table 2: Performance Evaluation of Different Models on Test Datasets for Bulgarian, Polish, Croatian, Slovene and Russian Language

4.4 Optimized Weight Selection

After fine-tuning several models, we combined them into an ensemble to achieve better results in different situations. To get the best performance from our ensemble, we tried out different ways of weighting each model's predictions. For every candidate set of weights, we looked at how well the combined predictions worked on our evaluation set by measuring F1 score, precision and recall. We then chose the weights that gave us the highest F1 score. The full optimization procedure is detailed in Algorithm 1. This method helps us make the most of each model's strengths and improves the overall accuracy of our final predictions.

4.5 Weighted Soft Voting Ensemble

After selecting the optimal set of weights, we applied these weights to combine the predictions from our three best fine-tuned models for each language, creating an ensemble model(X=1). We have used the best three fine-tuned model (e_m1, e_m2 and e_m3) for ensemble according to their F1 score. This process was repeated for all four languages. By using the optimized weights, the ensemble model leverages the strengths of each individual model.

Algorithm 1 Soft Voting Ensemble Weight Optimization

Require: Set of base model predictions $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n\}$, true labels $\mathbf{y}_{\text{true}}$, candidate weight sets $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ where $\sum_{i=1}^{k} w_i = 1$

Ensure: Optimal weight vector $\mathbf{w}^*$ maximizing F1 score on evaluation set

Initialize $F1_{\text{best}} \leftarrow 0$
Initialize $\mathbf{w}^* \leftarrow$ arbitrary initial weights
for $(j, \mathbf{w})$ in enumerate($\mathcal{W}$) **do**
  $\mathbf{S} \leftarrow \sum_{i=1}^{n} w_i \mathbf{P}_i$ {Compute weighted ensemble probabilities}
  $\hat{\mathbf{y}} \leftarrow \arg\max_c \mathbf{S}$ {Predict class with highest probability}
  Compute $F1_{\text{current}} \leftarrow \text{F1}(\mathbf{y}_{\text{true}}, \hat{\mathbf{y}})$
  if $F1_{\text{current}} > F1_{\text{best}}$ **then**
    $F1_{\text{best}} \leftarrow F1_{\text{current}}$
    $\mathbf{w}^* \leftarrow \mathbf{w}$
  end if
end for

return $\mathbf{w}^*, F1_{\text{best}}$

5 Results and Analysis

In this section, we have explained a detailed comparison of performance results for both single and ensemble models across different Slavic languages.

5.1 Parameter Setting

For our experiments on Slavic language classification, we carefully selected hyperparameters based on preliminary grid searches using optuna(Akiba et al., 2019) and best practices from related works. For learning rate, we tested values from 2×10^{-4} to 2×10^{-3}, choosing the best per language based on validation accuracy and convergence. Weight decay was set between 0.01 and 0.02 to control overfitting, as higher values caused underfitting. Table 4 shows parameter settings for different models.

In Table 4, *lr*, *optim*, *w_d* and *e_s* represents *learning_rate*,*optimizer*, *weight_decay* and *Early_Stopping* and respectively.

Language	Model	Weight	Acc	Precision	Recall	F1
Bulgarian	BERTić*(Ljubešić and Lauc, 2021)	NA	0.872	0.825	0.961	0.888
	xlm-roberta-large(Conneau et al., 2019)	NA	0.816	0.5357	0.833	0.652
	bert-web-bg(Marinova et al., 2023)	NA	0.759	0.454	0.833	0.588
	Ensemble *(Combining upper 3 models)*	[0.6, 0.2, 0.2]	0.879	0.837	0.958	0.894
Polish	herbert(Mroczkowski et al., 2021)	NA	0.843	0.861	0.899	0.88
	Polbert(Kłeczek, 2020)	NA	0.791	0.713	0.882	0.788
	Polish-roberta(Semary et al., 2023)	NA	0.884	0.907	0.918	0.912
	Ensemble *(Combining upper 3 models)*	[0.4, 0.1, 0.5]	0.902	0.917	0.943	0.930
Slovene	SloBERTa (Ulčar and Robnik-Šikonja, 2021)	NA	0.801	0.783	0.821	0.801
	CroSloEngual BERT(Ulčar and Robnik-Šikonja, 2020)	NA	0.75	0.803	0.692	0.743
	SlovakBERT(Pikuliak et al., 2021)	NA	0.867	0.719	0.962	0.823
	Ensemble (Combining upper 3 models)	[0.4, 0.1, 0.5]	0.864	0.729	0.924	0.815
Russian	Conversational RuBERT (Galimzianov and Vyshegorodtsev, 2024)	NA	0.69	0.844	0.721	0.778
	RuBERT-tiny(Dale, 2022)	NA	0.633	0.691	0.672	0.681
	ruBert-base(Zmitrovich et al., 2023)	NA	0.591	0.643	0.612	0.6273
	Ensemble *(Combining upper 3 models)*	[0.5, 0.25, 0.25]	0.714	0.83	0.779	0.804

Table 3: Performance Evaluation of Different Models on Evaluation Dataset for Bulgarian, Polish, Croatian, Slovene and Russian Languages

Language	lr	optim	w_d	e_s
Bulgarian	$3e^{-4}$	Paged Adamw	0.02	2
Polish	$2e^{-4}$	Paged Adamw	0.02	2
Slovene	$2e^{-3}$	Adam	0.01	2
Russian	$2e^{-3}$	Adam	0.01	2

Table 4: Parameter settings for different models

5.2 Evaluation Metrics

The performance of various models has been evaluated using precision, recall, and F1 metrics.

5.3 Comparative Analysis

Table 2 and table 3 presents a comparative evaluation of various transformer-based models and their ensemble combinations across Bulgarian, Polish, Slovene, and Russian languages.

For Bulgarian, the BERTić model achieved strong results on the test set with an F1-score of 0.861. But the ensemble approach combining BERTić, xlm-roberta-large, and bert-web-bg outperformed individual models, achieving the highest F1-score of 0.867. In the case of Polish, also the ensemble of three single models achieved the best performance, reaching an F1-score of 0.902 on the test set. But in case of the Slovene language, we have seen that the single model SlovakBERT outperformed the

ensemble approach with an F1 score of 0.823. Finally, for Russian combining three models in an ensemble led to better results than any single model. The ensemble approach achieved a solid F1-score of 0.804.

5.4 Error Analysis

During evaluation, we observed that sentences containing multiple persuasion techniques are sometimes misclassified by the model. Although the use of ensemble modeling has reduced this issue, it still persists. Additionally, the presence of neutral sentences that do not show clear persuasive intent but may subtly influence the audience makes the classification process more challenging. Such neutral sentences are often inconsistently classified, with the model sometimes labeling them as persuasive and other times not.

Example Text:"Poseł Zbigniew Bogucki: . . . to dlaczego ma to być w przypadku prezydencji. (Oklaski)" *[MP Zbigniew Bogucki: . . . so why should it be the case in the situation of the presidency. (Applause)]*

Expected Prediction: True

Model Prediction: False

This example highlights the challenge subtle persuasive strategies like rhetorical questions may be missed by the model, while audience cues like applause cannot be used as the sole basis for classification.

6 Conclusion

This research explored the detection of persuasion techniques in Slavic languages using both single transformer models and weighted ensemble approaches. After applying preprocessing and fine-tuning language-specific models, our results show that ensemble methods generally perform better than individual models for most languages. For Slovene, the single SlovakBERT model slightly outperformed the ensemble, though the improvement was minimal. However, most of these results are close to the baseline. Nevertheless, it is insightful that even without a large dataset or external data, it is possible to achieve satisfactory results with limited data. This work contributes valuable insights and practical solutions for combating misinformation and manipulation in multilingual digital environments.

7 Limitations & Future Work

This study have a relatively small size of the training data, especially for languages like Slovene. Additionally, the complexity and length of many sentences in the dataset make accurate detection more challenging. For future work we want to explore techniques like data augmentation techniques and semi-supervised learning could help improve model robustness in low-resource settings. We also aim to experiment with more advanced transformer architectures and state-of-the-art large language models (LLMs).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

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.

David Dale. 2022. . [Online; posted 12-June-2022].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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. *arXiv preprint arXiv:2105.09284*.

Dmitrii Galimzianov and Viacheslav Vyshegorodtsev. 2024. Conversational rubert for detecting competitive interruptions in asr-transcribed dialogues. *arXiv preprint arXiv:2407.14940*.

Omid Ghahroodi and Ehsaneddin Asgari. 2024. Hierarchyeverywhere at semeval-2024 task 4: Detection of persuasion techniques in memes using hierarchical text classifier. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1727–1732.

Dariusz Kłeczek. 2020. Polbert: Attacking polish nlp tasks with transformers. In *Proceedings of the PolEval 2020 workshop*, pages 79–88.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Iva Marinova, Kiril Simov, and Petya Osenova. 2023. Transformer-based language models for Bulgarian. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 712–720, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Robert Mroczkowski, Piotr Rybak, Alina Wr ’oblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Nikolaos Nikolaidis, Nicolas Stefanovitch, and Jakub Piskorski. 2023. On experiments of detecting persuasion techniques in polish and russian online news: Preliminary study. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 155–164.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2021. Slovakbert: Slovak masked language model.

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.

Denise Scannell, Linda Desens, Marie Guadagno, Yolande Tra, Emily Acker, Kate Sheridan, Margo Rosner, Jennifer Mathieu, and Mike Fulk. 2021. Covid-19 vaccine discourse on twitter: A content analysis of persuasion techniques, sentiment and mis/disinformation. *Journal of health communication*, 26(7):443–459.

Noura A Semary, Wesam Ahmed, Khalid Amin, Paweł Pławiak, and Mohamed Hammad. 2023. Improving sentiment classification using a roberta-based hybrid model. *Frontiers in human neuroscience*, 17:1292010.

Cecilie S Traberg, Trisha Harjani, Jon Roozenbeek, and Sander Van Der Linden. 2024. The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports*, 14(1):4205.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Sloberta: Slovene monolingual large pretrained masked language model. *Proceedings of Data Mining and Data Warehousing, SiKDD*, pages 17–20.

M. Ulčar and M. Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.