Robust Detection of Persuasion Techniques in Slavic Languages via Multitask Debiasing and Walking Embeddings

Ewelina Księżniak and Krzysztof Węcel and Marcin Sawiński
ewelina.ksiezniak, krzysztof.wecel, marcin.sawinski@ue.poznan.pl
Poznań University of Economics and Business
Poznań, Poland

Abstract

Our approach to Subtask 1 integrates fine-tuned multilingual transformer models with two complementary robustness-oriented strategies: Walking Embeddings and Content-Debiasing. With the first, we tried to understand the change in embeddings when various manipulation techniques were applied. The latter leverages a supervised contrastive objective over semantically equivalent yet stylistically divergent text pairs, generated via GPT-4. We conduct extensive experiments, including 5-fold cross-validation and out-of-domain evaluation, and explore the impact of contrastive loss weighting.

1 Introduction

This paper presents our solution to Subtask 1 of the Shared Task on the Detection and Classification of Persuasion Techniques in Texts for Slavic Languages. The task focuses on identifying whether specified text fragments contain any persuasive techniques, according to a predefined taxonomy. Training data was available in four Slavic languages—Polish, Slovenian, Bulgarian, and Russian—while the test set also included Croatian. A detailed overview of the datasets is provided in Piskorski et al. (2025).

Our approach combines standard fine-tuning of transformer-based models with two complementary techniques designed to improve robustness. The first, Walking Embeddings, analyzes how sentence representations evolve as words are incrementally added. The second, Content-Debiasing, introduces a multitask setup with a contrastive learning objective, leveraging pairs of semantically equivalent texts—one with and one without persuasive language—to help the model disentangle content from stylistic features.

2 Related Work

Persuasion detection has gained significant attention in NLP, particularly in connection with the fine-grained identification of rhetorical strategies and propaganda techniques. SemEval-2020 Task 11 (Da San Martino et al., 2020) formalized the task as both binary classification (persuasive vs. non-persuasive) and span-level classification into specific techniques, such as *Appeal to Fear*, *Loaded Language*, or *Name Calling*. Transformer-based models, especially RoBERTa and BERT variants, have been widely adopted for this task, often enhanced with additional features or ensemble methods. For example, the top-ranked systems in SemEval-2020 and 2021 used ensembles of RoBERTa and domain-adapted BERT models, sometimes combined with task-specific layers or external lexicons to improve detection of subtle rhetorical signals (Dimitrov et al., 2021). Similarly, in the CLEF-2024 CheckThat! Lab Task 3, participating teams applied fine-tuning of BERT-based models, including techniques such as data augmentation with word alignment to project labels from source texts onto machine-translated target texts (Piskorski et al., 2024).

Model debiasing aim to improve model robustness by reducing reliance on spurious correlations or stylistic artifacts in the input. In the context of NLP, debiasing has been applied to mitigate among others gender, racial, and stylistic biases in representations and predictions (Zhao et al., 2018; Liang et al., 2020). A common strategy is to introduce auxiliary objectives that penalize the model when it relies on confounding factors rather than semantically meaningful content. One possible method is contrastive learning, which encourages similar representations for semantically equivalent inputs while pushing apart dissimilar ones (Gunel et al., 2021). In NLP, contrastive objectives are often applied over paraphrase pairs, style-transferred sen-

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 224–230

tences, or counterfactual augmentations, helping models to align content representations across superficial differences. This has proven especially effective in tasks like sentiment analysis, sarcasm detection (Jia et al., 2024), and social bias mitigation, where the boundary between content and tone is particularly subtle.

Multitask learning is a training paradigm in which a model learns to perform multiple related tasks simultaneously, often leading to better generalization and robustness across domains (Caruana, 1997). By sharing representations between tasks, the model can leverage auxiliary signals to improve the performance of the main objective.

3 System Description

3.1 Model debiasing

To enhance robustness and reduce overfitting to superficial persuasive cues, we implemented a content-debiasing mechanism based on multitask learning with supervised contrastive loss. The goal was to help the model disentangle semantic content from stylistic elements associated with persuasion. Our method is an adaptation of the approach proposed by Jia et al. (2024), who applied topic debiasing via contrastive learning in the context of multimodal sarcasm detection, combining textual and visual signals.

For each training example, we automatically generated a pair of texts using the GPT-4o API with a temperature setting of 0.2 to ensure controlled outputs (OpenAI, 2025). The original text contained annotated spans of persuasive language, while the rewritten version preserved the meaning but neutralized the style within those spans. To guide generation, we used the following prompt:

> *You will be given a text that contains one or more marked spans. Each span is marked like this: [start span=TECHNIQUE]... [end span]. Your task is to rewrite **only** the text inside each span to make it **neutral and objective**, removing the influence of the persuasive technique given in the tag. Keep the language and structure of the original text outside the span untouched.*
>
> *Example:*
>
> *Original: Ludzie [start span=AppealToFear]umrą, jeśli nie zrobimy tego teraz![end span] To nasza jedyna szansa.*
>
> *Neutralized: Ludzie [start span=AppealToFear]są zaniepokojeni możliwymi konsekwencjami dalszego zwlekania.[end span] To nasza jedyna szansa.*

Span annotations were sourced from Subtask 2 and directly referenced in the prompt.

The resulting pairs were used in a multitask setup: the primary task was binary classification (detecting the presence of any persuasive technique), and the auxiliary task employed a supervised contrastive objective. For auxiliary task: both original and neutralized texts were encoded using a shared XLM-RoBERTa-base model (Conneau et al., 2020), and their `[CLS]` embeddings were used to compute Supervised Contrastive Loss (SupConLoss), which encourages representations of similar (e.g., semantically aligned) inputs to be pulled closer while pushing apart dissimilar ones within a supervised setup (Khosla et al., 2020). Despite semantic equivalence, the pairs were labeled as negatives, as they differed stylistically. Pairwise similarities between embeddings were computed using cosine similarity over normalized vectors, scaled by a temperature parameter. The resulting similarity matrix served as the foundation for the contrastive loss, which penalized the model when stylistically divergent pairs were embedded too closely.

The total loss combined cross-entropy (for classification) and contrastive loss, weighted by a tunable hyperparameter λ. We set $\lambda = 0.3$ in our submission experiments, balancing the influence of both objectives. We trained three model variants on distinct training splits, each selected using a different random seed. All models were fine-tuned with a learning rate of $1e{-}5$, batch size of 16, and a maximum of 10,000 steps. Early stopping was applied with a patience of 2. Further analysis of these choices is presented in the Experiments section.

3.2 Walking embeddings

In this approach the final classification method is based on logistic regression applied to sentence embeddings, optionally extended with embeddings of sentence halves to better capture rhetorical structure. We employed multilingual embedding models (Jina (Sturua et al., 2024) and E5 (Wang et al., 2024)) to generate vector representations of text fragments. This approach enables the model to differentiate between neutral and persuasive content by capturing semantic trajectories within sentences. All sentences were encoded individually, and their embeddings were used directly for classification. Further implementation details and evaluation results are presented in Experiments Section.

Our choice of logistic regression (LR) was motivated by its close functional similarity to the softmax classification head commonly

used in transformer-based models such as `BERTForSequenceClassification`. Both methods operate on top of fixed-length embedding vectors and serve as simple, interpretable models for binary or multiclass classification. In our case, LR serves as a lightweight yet effective classifier that allows us to focus on the properties of the embeddings themselves, rather than the complexity of the classification model. This aligns with our study's goal of analyzing how well rhetorical anomalies can be captured through embedding space characteristics.

As for the embedding models, we selected Jina and E5 based on recent benchmark results. Both have demonstrated strong results on a variety of sentence-level tasks while maintaining relatively low computational requirements. This made them well-suited for local execution, which was a practical consideration for our study. We prioritized models that enabled rapid experimentation and interpretability without relying on large-scale infrastructure.

4 Results on test

Table 1 presents the results obtained using the described methods on the test set. For Croatian, the highest performance was achieved with walking embeddings method, while for all other languages, the content debiasing approach yielded superior results. According to the official ranking, our system achieved first place for Croatian and Bulgarian, second place for Slovenian, third place for Polish and fourth place for Russian. Detailed analysis and additional findings are provided in the accompanying report (Piskorski et al., 2025).

Lang.	BG	HR	PL	RU	SI
Acc.	86.11	95.95	86.97	75.76	89.32
Prec.	83.37	96.97	86.48	83.67	77.20
Rec.	92.79	94.12	94.16	84.23	94.90
F1	87.83	95.52	90.16	83.95	85.14

Table 1: Performance of the FactUE team per language and run on Subtask 1. For Croatian (HR), the results correspond to a logistic regression model using JinaEmbeddings as described in 5.2. For all other languages, results are obtained using the debiasing approach with $\lambda = 0.3$ as described in 5.1.

5 Experiments

5.1 Model debiasing

In the first step, to establish a baseline, we fine-tuned two multilingual transformer models: mDeBERTa-v3-base (Microsoft, 2023) and XLM-RoBERTa (Conneau et al., 2020). To explore optimal training dynamics, we experimented with several learning rates: 5e-6, 2e-6, 1e-5, 2e-5, and 3e-5. Each configuration was trained three times using different random seeds (42, 100, 1111). Based on overall performance across these runs, we selected a fixed learning rate of 1e-5 for subsequent experiments. For the construction of the training, validation, and test sets, we combined and shuffled the datasets labeled as train and trial, which were provided by the organizers.

To evaluate our proposed content-debiasing method under limited data conditions, we conducted 5-fold cross-validation, assessing results separately for each language. Additionally, to measure the robustness of the model—our method's primary goal—we evaluated it on an out-of-domain test set: a sample from the English binary persuasion classification dataset released as part of SemEval 2020, which consisted of 3,186 annotated examples. Due to time and resource constraints prior to the submission deadline, we were only able to test the model's behavior for a limited set of lambda values: 0.1, 0.2, and 0.3. Based on these preliminary results, we selected lambda equal 0.3 for the final submission model. However, following the submission, we conducted additional experiments exploring a broader range of lambda values to better understand the method's sensitivity and performance across different regularization strengths.

Performance across different values of the contrastive loss weight (lambda) is summarized in Table 2. The table reports the average F_1 score for the positive class (*f1 pos*), computed via 5-fold cross-validation. The cross-validation was performed on a dataset created by merging the `train` and `trial` splits provided by the organizers. A lambda of 0 corresponds to the baseline (standard fine-tuning), while lambda 1 assigns equal weight to the primary and auxiliary tasks.

5.2 Walking embeddings

Walking embeddings is our original idea stemming from our other works on representation of text fragments. We observed that: change of order of words

λ	BG	PL	RU	SI	EN (OOD)
0.0	0.97	0.89	0.73	0.74	0.17
0.1	0.98	0.88	0.72	0.81	0.16
0.2	0.97	0.88	0.75	0.70	0.17
0.3	0.97	0.88	0.77	0.77	0.20
0.4	0.98	0.88	0.74	0.73	0.17
0.5	0.98	0.89	0.76	0.73	0.18
0.6	0.97	0.89	0.73	0.73	0.17
0.7	0.97	0.90	0.79	0.83	0.21
0.8	0.97	0.87	0.79	0.77	0.19
0.9	0.97	0.90	0.76	0.67	0.21
1.0	0.98	0.91	0.76	0.77	0.20

Table 2: Mean F1 scores per language for different values of contrastive loss weight λ. EN refers to the out-of-domain English test set.

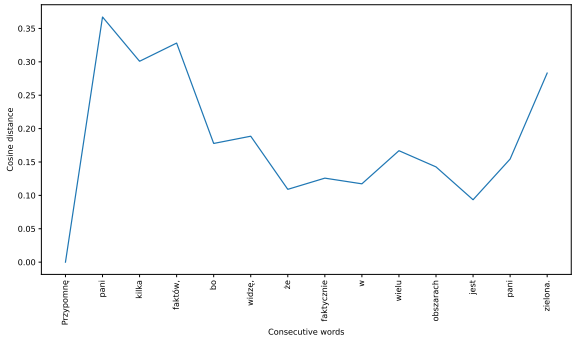

Figure 1: Cosine distances between E5 embeddings of the growing fragments of a sample sentence

(like for keywords) results in significant change of embeddings; encoding longer fragments does not allow to find a matching subsequence based only on embeddings.

In this approach we study the changes in the embeddings while new words are added. The assumption was that final classification of the sentence depends not only on the embedding of the whole sentence but it is also important what where the embeddings 'on the way'. Several experiments have been conducted.

In the first experiment we studied the change in distance, when new word was added to a sentence. Considering the example sentence from the training dataset: "Przypomnę pani kilka faktów, bo widzę, że faktycznie w wielu obszarach jest pani zielona", we built the following fragments: "Przypomnę", "Przypomnę pani", "Przypomnę pani kilka" and so on. Figure 1 presents the cosine distances between embeddings of consecutive growing fragments of a sentence. For the sample sentence, we were particularly interested in the distance between the last two fragments, because the last word, "zielona", was tagged as *Name_Calling-Labeling*. The end of the sentence "you are green" can be interpreted in different ways: label for somebody who know little or nothing, or referring to an ecologist. The sentence could end with phrase "you are an expert", and that should not be annotated by the system.

Unfortunately, we did not observe any specific change in embeddings where different end words were attempted, e.g., "green", "red", "expert", etc. Green seemed just as good as some other designations of a person. We also repeated the same chart for all sentences in the training dataset (figure 2). Green lines represent sentences labeled as 'false', and red – sentences with persuasion techniques. Here, we can observe that neutral sentences are positioned a little bit higher regarding semantic distance between fragments.

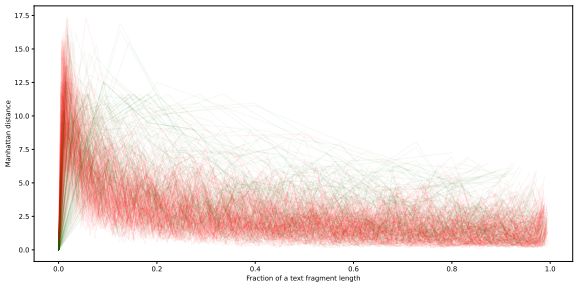

Figure 2: Manhattan distances between E5 embeddings of the growing fragments of all sentences in a training dataset

While reducing embeddings to pairwise distances offers an easy-to-understand perspective, it may overlook important structural nuances. To gain deeper insight, we also analyzed the trajectories of embeddings in their original high-dimensional space. For visualization purposes, we projected the embeddings onto two dimensions using Principal Component Analysis (PCA).

Figure 3 demonstrates our walking embeddings. The green arrow denotes the beginning of a sentence (i.e., the embedding of the first word). The red square represents the embedding of a manipulated sentences, while the blue square corresponds to a neutral sentences. Due to the standardized nature of public speaking, many sentences begin in similar regions of the embedding space.

Notably, the embeddings of manipulated ("red") sentences tend to be distinguishable from those of neutral ("blue") ones, which motivated their use in our classification task. However, we need to be careful in interpreting these visualizations, as dimensionality reduction techniques like PCA do not

fully preserve the complex relationships present in the original high-dimensional space. These visualizations serve only as a simplified aid to understanding the underlying patterns. Indeed, PCA applied to dense embedding vectors typically captures only a limited portion of the total variance.

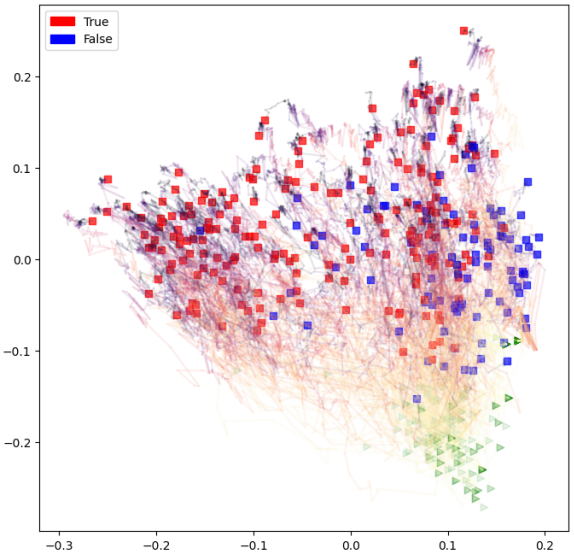

Figure 3: Traces of walking embeddings (E5) for all sentences with true/false classification

The final classification was performed using logistic regression applied to stacked two embeddings: those derived from the entire input text fragments and those obtained from their respective halves, to account for rhetorical structure. Our experiments involved analysis of complete rhetorical trajectories. However, due to time constraints, we did not develop a method to exploit the insights illustrated in Figure 2, leaving this as a direction for future work.

A key challenge is to identify rhetorical breaking points – positions in the text where the rhetorical flow deviates from expected patterns. For instance, consider a text fragment consisting of two consecutive sentences. Typically, the second sentence maintains coherence with the first, a property exploited by many training objectives such as next sentence prediction (NSP). However, in some cases, the second sentence may be unrelated, introduce unsupported conclusions, or shift the topic unexpectedly. Our proposed simplification is as follows: analyze the first sentence fragment, and if the subsequent sentence introduces an unexpected rhetorical shift, the model should be able to detect this as an anomaly.

We attempted two embedding models: Jina (jinaai/jina-embeddings-v3) (Sturua et al., 2024) and E5 (intfloat/multilingual-e5-large) (Wang et al., 2024). For our separated test dataset, F1 macro avg for Polish was 0.84 using Jina on single embeddings, and 0.85 on combined. For E5, we achieved 0.87 in both variants. Logistic regression performed better than XGB, which achieved 0.77 compared to 0.84 on the same input. Final submission was prepared by logistic regression trained on samples in all languages using extended embeddings. The models combining full and half-sentence embeddings returned better results than models using only full embeddings.

6 Conclusions

Our experiments confirm that contrastive content-debiasing improves model robustness across Slavic languages and leads to better generalization on out-of-domain data, including English. Cross-validation results show consistent gains in F1 score when supervised contrastive loss is used alongside standard fine-tuning. While the walking embeddings approach did not yield clearly discriminative patterns in embedding space, preliminary analyses suggest it may provide a useful lens for exploring how rhetorical structure evolves within sentences. Although our experiments did not reveal consistent accumulation of persuasive cues, the observed embedding trajectories highlight areas for further investigation, particularly in identifying rhetorical shift points. These findings suggest that stylistic regularization and embedding dynamics can be complementary tools for enhancing persuasion detection systems.

References

Rich Caruana. 1997. Multitask learning. In *Learning to learn*, pages 95–133. Springer.

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. `https://huggingface.co/xlm-roberta-base`. *Preprint*, arXiv:1911.02116. Accessed in May 2025.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Preslav Nakov, and James Glass. 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), Spain. International Committee for Computational Linguistics.

Dimitar Dimitrov, Preslav Nakov, Giovanni Da San Martino, Alberto Barrón-Cedeño, Bilyana Taneva, Wajdi Zaghouani, Momchil Hardalov, and Henning Wachsmuth. 2021. Semeval-2021 task 6: Detection of persuasive 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.

Beliz Gunel, Canwen Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations (ICLR)*.

Mengzhao Jia, Can Xie, and Liqiang Jing. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, Vancouver, Canada. AAAI Press. Article ID: 29795.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Paul Pu Liang, Thomas Manzini, Ryan Shelby, Sumeet Singh, Rahul Jha, Carson Schwemmer, Roi Reichart, Jonathan Zittrain, Jennifer Hutson, Dan Jurafsky, and 1 others. 2020. Towards understanding and mitigating social biases in language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5504–5515, Online. Association for Computational Linguistics.

Microsoft. 2023. mdeberta-v3-base. `https://huggingface.co/microsoft/mdeberta-v3-base`. Accessed in May 2025.

OpenAI. 2025. Gpt-4o api. `https://platform.openai.com/docs/models/gpt-4o`. Accessed in May 2025.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. 2024. Overview of the clef-2024 checkthat! lab task 3: Multilingual detection of persuasion techniques in texts. In *Proceedings of the 15th Conference and Labs of the Evaluation Forum (CLEF 2024)*, Grenoble, France. CEUR Workshop Proceedings. To appear.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, 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.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix - Experimental Details for model debiasing approach

To ensure clarity and reproducibility, we provide the following detailed description of our experimental setup.

A.1 Model and Tokenizer - model debiasing approach

We used the `xlm-roberta-base` model along with its associated tokenizer, loaded via Hugging Face's.

A.2 Hyperparameters - model debiasing approach

We used the following training configuration:

- Learning rate: `1e-5`
- Weight decay: `0.05`
- Batch size (train/eval): `16`
- Maximum training steps: `10,000`
- Evaluation frequency: every `100` steps
- Model saving frequency: every `100` steps (best model retained)
- Early stopping: patience of `2` evaluations
- Mixed precision (FP16): enabled
- Maximum sequence length: `128`
- Optimization objective: F1 score of the positive class (`f1_pos`)

A.3 Random Seed and Reproducibility

We fixed the random seed to `42` across all components, including data splits and model initialization. The CUDA device was set via `CUDA_VISIBLE_DEVICES`. All models were trained using PyTorch and Hugging Face Transformers.

A.4 Out-of-Domain Evaluation Sample

For the out-of-domain evaluation, we used a dataset released as part of the CLEF 2024 CheckThat! Lab, specifically from the adversarial persuasion detection subtask. The dataset consisted of 3,186 English-language examples and was originally sourced from the SemEval 2020 Task 6 binary persuasion classification dataset, where it served as the development split.

- **Current usage:** CLEF 2024 CheckThat! Lab – adversarial persuasion detection task – dev split.
- **Original source:** SemEval 2020 Task 6 (Zampieri et al., 2020).
- **Language:** English.
- **Sample size:** 3,186 examples.
- **Annotation schema:** Each instance is annotated with a binary label indicating whether the text is persuasive or non-persuasive.
- **Evaluation role:** This dataset was used strictly for out-of-domain evaluation. It was not used during training or model selection.