RDproj at SemEval-2024 Task 4: An Ensemble Learning Approach for Multilingual Detection of Persuasion Techniques in Memes

Yuhang Zhu
Uppsala University / Uppsala, Sweden
yuhang.zhu.2485@student.uu.se

Abstract

This paper introduces our bagging-based ensemble learning approach for the SemEval-2024 Task 4 Subtask 1, focusing on multilingual persuasion detection within meme texts. This task aims to identify persuasion techniques employed within meme texts, which is a hierarchical multilabel classification task. The given text may apply multiple techniques, and persuasion techniques have a hierarchical structure. However, only a few prior persuasion detection systems have utilized the hierarchical structure of persuasion techniques. In that case, we designed a multilingual bagging-based ensemble approach, incorporating a soft voting ensemble strategy to effectively exploit persuasion techniques' hierarchical structure. Our methodology achieved the second position in Bulgarian and North Macedonian, fifth in Arabic, and eleventh in English.

1 Introduction

Memes have gained immense popularity among the younger generation due to their entertaining nature. However, some memes can lead teenagers towards extreme ideas by employing persuasion techniques. Even well-educated people often need help to identify misleading memes. Thus, the development of a persuasion detection system holds significant value. This study aims to create a system to identify persuasion techniques within meme texts. This task is a multilabel and hierarchical classification task since memes may contain multiple persuasion techniques, and techniques have hierarchical structure (Dimitrov et al., 2024).

A description of the corpus provided by SemEval-2024 Task 4 (Dimitrov et al., 2024) reveals significant imbalances in the training data for the techniques. For instance, while there are 1990 instances for the "Smears" technique, only 258 instances pertain to "Whataboutism." Moreover, the training data for each technique is smaller compared with the entire corpus, leading to the imbalance between positive and negative instances for each technique. These observations lead us to formulate the following research questions: 1) How can we mitigate the data imbalance between techniques? 2) How can we ease the imbalance between positive and negative instances for each technique? 3) How can we effectively leverage the hierarchical structure of techniques? We devise a bagging-based ensemble learning system employing a soft voting strategy to solve these questions. We group techniques into ten subsets based on the amount of their training data and the hierarchical structure (Dimitrov et al., 2024), and construct a training set for each subset. Subsequently, we train classifiers (base learners), XLM-RoBerta$_{large}$[1] models with a classifier head, on these training sets. Finally, we compute the final distribution through a weighted average of the probability generated by classifiers, with a model of identical structure generating the weights in this step.

While our approach attained the second position in Bulgarian and North Macedonian, fifth in Arabic, and eleventh in English, the performance of our weight model did not exhibit significant improvement compared to our baseline. Moreover, the lower-resource techniques continue to suffer from imbalances between positive and negative instances. Our code is publicly available at https://github.com/Yuhang-Zhu-nlp/semeval2024_RDproj.

2 Background

2.1 Persuasion Detection

Previous research on persuasion detection has explored traditional classification techniques across a range of domains. Regarding data augmentation, Modzelewski et al. (2023) experimented with

[1] https://huggingface.co/FacebookAI/xlm-roberta-large

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 181–187
June 20-21, 2024 ©2024 Association for Computational Linguistics

enhancing performance by expanding the training set using the DeepL API to translate data from source languages to target languages. Similarly, Falk et al. (2023) introduced a data augmentation method based on back-translation in the same year. Regarding text representation, Qachfar and Verma (2023) proposed a technique to generate language-agnostic features specific to this task, which were then concatenated with the CLS representation provided by XLM-RoBERTa to generate the final representation. Ensemble learning has also been explored in this domain. Purificato and Navigli (2023) developed a multilingual bagging-based ensemble learning system, combining five different BERT models using a soft voting strategy. Because of BERT's exceptional performance in sentence classification tasks, it has become a cornerstone in recent research, with almost all contemporary studies incorporating BERT into their methodologies (Costa et al., 2023; Ojo et al., 2023).

2.2 Ensemble Learning

The term ensemble learning is basically to improve the model's performance by combining different models (base learners) (Dong et al., 2020). Presently, ensemble learning strategies primarily include bagging, boosting, and stacking. Among these, bagging is training models on distinct datasets and combining them. One of the most renowned bagging-based ensemble learning algorithms is random forest (Cutler et al., 2012), which trains numerous decision trees on different data subsets and then combines these trees using a voting strategy. Regarding voting strategies, there are two main approaches: hard voting (Mohamed Kamr and Mohamed, 2022) and soft voting (Purificato and Navigli, 2023). Soft voting generates the final distribution by computing the weighted average of distributions from base learners, and has become a prevalent strategy in classification tasks (Xu et al., 2016; Kumari et al., 2021). Purificato and Navigli (2023) devised a bagging-based multilingual ensemble learning approach, employing five different BERT models with a soft voting strategy in this task. Their approach secured the first position in English during SemEval 2023, underscoring the effectiveness of bagging-based ensemble learning in this context. However, their approach determined model weights based on the normalized F1-micro score of diverse BERT models, ignoring the potential variability in model performance across different techniques.

2.3 Data

We use both the corpus offered by SemEval-2024 Task 4 Subtask 1 (Dimitrov et al., 2024) which contains English text of memes with 20 persuasion techniques and the corpus provided by SemEval-2023 Task 3 Subtask 3 (Piskorski et al., 2023) which includes news articles in six languages, English, German, French, Russian, Polish, and Italian, with 23 techniques.

3 System Overview

3.1 Data Preprocessing

In this task, we only focus on 20 techniques, but the corpus provided by SemEval-2023 Task 3 Subtask 3 contains 23 techniques. In that case, We have simply removed the three extra techniques from the label set of each data. The corpus provided by SemEval-2024 Task 4 Subtask 1 includes lots of meaningless symbols like "\n", we just simply remove them from the text. Moreover, we lowercase all data of both corpora.

3.2 Technique Grouping

To utilize the hierarchical structure of techniques, we categorize them into seven subsets based on their hierarchical structure (Dimitrov et al., 2024). For each subset, we assess whether data imbalance exists among the techniques. If imbalances exist, we create new subsets and copy the affected techniques or divide the subset into smaller subsets. For example, in the initial grouping, "Loaded Language", "Exaggeration/Minimisation", "Flag-waving", and "Appeal to fear/prejudice" are grouped in a subset. However, the training data for "Loaded Language" significantly outnumbers those for the other three techniques, so we separate "Loaded Language" into a new subset while removing it from the original subset. Additionally, if some techniques unavoidably suffer from data imbalances, we copy them to a new subset (supporting subset). Through this process, we ultimately establish ten distinct subsets, and the results of grouping are shown in Appendix.

3.3 Corpus Creating

For each technique subset, we first sample all data in the corpus provided by Semeval-2023 Task 3 Subtask 3 (Piskorski et al., 2023) (in the following section, we call it positive data). Then, we sample the data without techniques in the subset (in the following section, we call it negative data). Next, we create the second corpus by doing the above step in the corpus offered by Semeval-2024 Task 4 Subtask 1 (Dimitrov et al., 2024).

3.4 Model Structure

We have 11 models in our approach, including 10 base learners and a weight model. All models have the same structure which is shown in Figure 1.

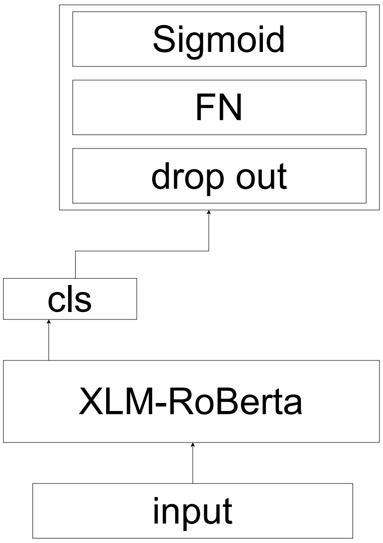

Figure 1: The structure of the base learners, and the weight model.

3.5 Training Strategy

Firstly, for each corpus sampled in the corpus provided by Semeval-2023 Task 3 Subtask 3 (Piskorski et al., 2023), we train a base learner on it (we call it pretrain in the following text). Then we fine-tune a base learner on each corpus sampled in the corpus offered by Semeval-2024 Task 4 Subtask 1 (Dimitrov et al., 2024). The task for base learners is to predict which persuasion techniques are applied in the given text. As for the weight model, we pretrain it on the original corpus provided by Semeval-2023 Task 3 Subtask 3 (Piskorski et al., 2023), and then fine-tune it on the corpus offered by Semeval-2024 Task 4 Subtask 1 (Dimitrov et al., 2024). The task of the weight model is to predict which technique subsets the persuasion techniques used in the given text belong to.

3.6 Prediction Pipeline

The prediction pipeline begins with preprocessing the text, which involves lowercasing and removing meaningless symbols. Subsequently, the text is sent to each base learner to obtain the technique distributions from each base learner. Similarly, the text is also sent to the weight model, and the output of the weight model is activated using softmax to generate the weight for soft voting. Finally, the final distribution is calculated using Equation (1).

$$D_{final} = \sum_{i=0}^{10} w_i D_i \tag{1}$$

where D_{final} is the final distribution, D_i is the distribution generated by the i^{th} base learner, and w_i is the weight generated by weight model for the i^{th} base learner.

4 Experimental Setup

We use binary cross-entropy (BCE) with weight as our loss function for each base learner. The equation is below:

$$\mathcal{L}(x_j, y_j) = \sum_{j=0}^{20} w_j(y_j log x_j - (1-y_j) log(1-x_j)) \tag{2}$$

where w_j is the weight for the j^{th} technique, y_j is the boolean value for the j^{th} technique, and x_j is the probability generated by the model for the j^{th} technique. We use BCE without weight for the weight model.

4.1 Training Setup

Each base learner has three hyperparameters: weights in the loss function, learning rate, and dropout rate. We set the learning rate to 2e-6 and the dropout rate to 0.2 for all base learners. The weights assigned to techniques belonging to the subset used to create the corpus on which the base learner is trained are set to 2, while all other techniques are assigned a weight of 1. Similarly, we use the same learning and dropout rates for the weight model as the base learner. During pretraining, we train each base learner for 60 epochs and the weight model for 50 epochs. During fine-tuning, we train each base learner for 20 epochs and the weight model for 10 epochs. The batch size is set to 16 for base learners and 8 for the weight model. we select 0.22 as our classification threshold.

4.2 Evaluation Metrics

Hierarchical-F1 (Kiritchenko et al., 2006) is used in this research. The benefit of the hierarchical-F1 is that it takes the hierarchical structure of techniques into account.

5 Results

5.1 Official Ranking

Table 1 shows our results in SemEval-2024 Task 4 Subtask 1. Although we get only the eleventh position in English, our results in three languages that are used to test zero-shot are competitive. We achieve the second position in both Bulgarian and North Macedonian, and the fifth position in Arabic.

5.2 Weight Model

We design a baseline model by removing the weight model, and set the weights for soft voting as $\frac{1}{10}$. In Table 2, we can find that our baseline and approach get almost the same score in English, Bulgarian, and North Macedonian. However, our baseline gets a relatively higher score in Arabic, which means that our weight model does not work well.

5.3 Error Analysis

In this section, we are aiming to find out the behaviour of our model facing different inputs by analyzing the samples which make our model give a wrong prediction in the dev set provided by SemEval-2024 Task 4 Subtask 1.

Text: IF YOU SAY WE'RE IN THE MIDDLE OF A DEADLY PANDEMIC BUT YOU STILL SUPPORT OPEN BORDERS\\n\\nYOU'RE EITHER A LIAR OR A COMPLETE MORON

Gold labels: Loaded Language, Name calling/Labeling, Black-and-white Fallacy/Dictatorship, Smears

Our prediction: Appeal to fear/prejudice, Black-and-white Fallacy/Dictatorship, Loaded Language, Name calling/Labeling, Smears

Weight vector: 0.0748, 0.0748, 0.0748, 0.1978, 0.2029, 0.0749, 0.0752, 0.0748, 0.0750, 0.0748

In this sample, we correctly identify all gold labels but detect "Appeal to fear/prejudice" by mistake. Analysis of the weight vector reveals that our weight model assigns a relatively higher weight of 0.2029 to the base learner trained on the corpora sampled for the subset (we call the base learner trained on the subset in the following text) containing "Appeal to fear/prejudice". However, it does not assign higher weights to subsets containing other techniques in the gold labels, except for "Loaded Language". To comprehend why our model can still make correct predictions despite the weight model's failure, we examine the output of several base learners. We observe that almost all base learners assign high probabilities to "Loaded Language", "Name calling/Labeling", and "Smears", indicating that each base learner can support techniques not included in the subsets on which they are trained. This suggests that each base learner can support the target techniques that are not included in the subsets they trained on.

Text: Name: Ted Bundy\\nVictims: 30\\n\\nName: Al Gore\\nVictims: ???

Gold labels: Reductio ad hitlerum, Smears

Our prediction: Name calling/Labeling

Weight vector: 0.1000, 0.1000, 0.1000, 0.1000, 0.1000, 0.1000, 0.1000, 0.1000, 0.1000, 0.1000, 0.1000

In this sample, we can find that our weight model does not work and give every subset a same weight. "Reductio ad hitlerum" is included in three technique subsets, and only the base learner trained on the supporting subset gives a high probability for this technique. However, other base learners give a very low probability, which shows our idea to create more subsets to support techniques suffering from data imbalance is working. The reasons for why we cannot distinguish "Reductio ad hitlerum" are 1) weight model cannot find which subsets the final prediction should be in, 2) positive and negative instances for "Reductio ad hitlerum" are too imbalanced, and our model tends to give a low probability.

Weight vector: IS THE BUNDY SHOOTOUT A FALSE FLAG?\\n

Gold labels: Doubt

Our prediction: Loaded Language, Name calling/Labeling, Doubt

Weight vector: 0.1663, 0.0958, 0.0922,

language	rank/nt	F1	T1F1
English	11/34	0.64288	0.75247
Bulgarian*	2/20	0.54089	0.56833
North Macedonian*	2/20	0.49869	0.51244
Arabic*	5/17	0.41129	0.47593

Table 1: The ranking of our approach in the official ranking of SemEval-2024 Task 4 Subtask 1. Languages with star are to test zero-shot. nt is the number of teams. F1 is the hierarchical-F1 score. T1F1 is the hierarchical-F1 score of the top-1 approach.

language	Our Model	Baseline
English	0.64288	0.64194
Bulgarian*	0.54089	0.54133
North Macedonian*	0.49869	0.49894
Arabic*	0.41129	0.41454

Table 2: The hierarchical-F1 score of our approach and the baseline on the test set.

> 0.0922, 0.0923, 0.0922, 0.0922, 0.0922,
> 0.0922, 0.0922

The weight model gives a higher weight for the first two subsets, which is correct because both subsets contain “Doubt”. Almost all base learners give a high probability for “Doubt”, which provide another evidence that base learners trained on other subsets can support gold labels. However, some base learners also give high probabilities for other two techniques in our prediction, resulting in wrong prediction. We should find a way to expand the gap between the weight of base learners trained on the subsets that include gold labels and on other subsets.

We can find some common elements in all samples. For example, “Loaded Language” and “Name calling/Labeling” are always predicted by mistake. A possible reason for this is that 0.22 is a reasonable threshold for some techniques but too small for some techniques which have rich training instances. Moreover, the accuracy of the weight model is not high enough.

6 Conclusion

In this study we build a persuasion detection system to distinguish which techniques are used in the given text of memes. Our system consists of ten base learners trained on different technique subsets and a weight model to generate the weight for soft voting. In the official ranking of SemEval-2024 Task 4 Subtask 1, we get competitive results in the zero-shot setting. However, our weight model does not work very well, and does not show a significant improvement compared with our baseline. The problems may be 1) the accuracy of the weight model is not high enough, 2) the gap between the weight of base learners trained on target subsets and other base learners is not big enough. Our idea to create a new technique subset to support techniques suffering from data imbalance seems feasible but the data imbalance between positive and negative instances of a technique is still a problem. The above discussion suggests the ideas to improve our approach. Firstly, we can improve the accuracy of the weight model by applying some new training techniques because our training method is very simple. Secondly, we need a more sophisticated technique grouping strategy which considers imbalance of positive and negative instances of a technique better.

Acknowledgements

I would like to express my appreciate for all participants who published their work in previous SemEval conferences. Additionally, I want to thank my teacher, Joakim Nivre, who gave me useful advice for my research.

References

Nelson Filipe Costa, Bryce Hamilton, and Leila Kosseim. 2023. CLaC at SemEval-2023 task 3: Language potluck RoBERTa detects online persuasion techniques in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1613–1618, Toronto, Canada. Association for Computational Linguistics.

Adele Cutler, D Richard Cutler, and John R Stevens. 2012. *Random forests*, pages 157–175. Springer.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International*

Workshop on Semantic Evaluation, SemEval 2024, Mexico City, Mexico.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258.

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.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19*, pages 395–406. Springer.

Saloni Kumari, Deepika Kumar, and Mamta Mittal. 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2:40–46.

Arkadiusz Modzelewski, Witold Sosnowski, Magdalena Wilczynska, and Adam Wierzbicki. 2023. DSHacker at SemEval-2023 task 3: Genres and persuasion techniques detection with multilingual data augmentation through machine translation and text generation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1582–1591, Toronto, Canada. Association for Computational Linguistics.

Abdulrahman Mohamed Kamr and Ensaf Mohamed. 2022. akaBERT at SemEval-2022 task 6: An ensemble transformer-based model for Arabic sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 885–890, Seattle, United States. Association for Computational Linguistics.

Olumide Ojo, Olaronke Adebanji, Hiram Calvo, Damian Dieke, Olumuyiwa Ojo, Seye Akinsanya, Tolulope Abiola, and Anna Feldman. 2023. Legend at ArAIEval shared task: Persuasion technique detection using a language-agnostic text representation model. In *Proceedings of ArabicNLP 2023*, pages 594–599, Singapore (Hybrid). Association for Computational Linguistics.

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

Fatima Zahra Qachfar and Rakesh Verma. 2023. ReDASPersuasion at SemEval-2023 task 3: Persuasion detection using multilingual transformers and language agnostic features. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2124–2132, Toronto, Canada. Association for Computational Linguistics.

Steven Xu, HuiZhi Liang, and Timothy Baldwin. 2016. UNIMELB at SemEval-2016 tasks 4A and 4B: An ensemble of neural networks and a Word2Vec based model for sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 183–189, San Diego, California. Association for Computational Linguistics.

A Grouping of Technique Labels

subset	techniques
Ethos_ad	Name calling/Labeling
	Doubt
	Smears
	Reductio ad hitlerum
	Whataboutism
Ethos_ad_s	Doubt
	Reductio ad hitlerum
	Whataboutism
Ethos_ot	Bandwagon
	Appeal to authority
	Glittering generalities (Virtue)
Pathos_m1	Loaded Language
Pathos_m2	Exaggeration/Minimisation
	Flag-waving
	Appeal to fear/prejudice
Logos_JU	Bandwagon
	Appeal to authority
	Flag-waving
	Appeal to fear/prejudice
	Slogans
Logos_ot	Repetition
	Obfuscation, Intentional vagueness, Confusion
Logos_DI	Whataboutism
	Misrepresentation of Someoneś Position (Straw Man)
	Presenting Irrelevant Data (Red Herring)
Logos_SI	Causal Oversimplification
	Black-and-white Fallacy/Dictatorship
	Thought-terminating cliché
support_imbalance	Bandwagon
	Reductio ad hitlerum
	Obfuscation, Intentional vagueness, Confusion

Table 3: Grouping of Technique Labels