CLaC at SemEval-2024 Task 4: Decoding Persuasion in Memes – An Ensemble of Language Models with Paraphrase Augmentation

Kota Shamanth Ramanath Nayak and **Leila Kosseim**
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
kotashamanthramanath.nayak@mail.concordia.ca,
leila.kosseim@concordia.ca

Abstract

This paper describes our approach to SemEval-2024 Task 4 subtask 1, focusing on hierarchical multi-label detection of persuasion techniques in meme texts. Our approach was based on fine-tuning individual language models (BERT, XLM-RoBERTa, and mBERT) and leveraging a mean-based ensemble model. Additional strategies included dataset augmentation through the TC dataset and paraphrase generation as well as the fine-tuning of individual classification thresholds for each class. During testing, our system outperformed the baseline in all languages except for Arabic, where no significant improvement was reached. Analysis of the results seem to indicate that our dataset augmentation strategy and per-class threshold fine-tuning may have introduced noise and exacerbated the dataset imbalance.

1 Introduction

The SemEval-2024 shared Task 4 (Dimitrov et al., 2024) proposed three distinct subtasks dedicated to identifying persuasion techniques conveyed by memes. The primary aim was to unravel how memes, integral to disinformation campaigns, employ various techniques to shape user perspectives. Subtask 1 focused on the analysis of textual content alone; while subtasks 2 and 3 involved the analysis of multimodal context that considers both textual and visual elements. Subtasks 1 and 2 used hierarchical multi-label classification metrics, while subtask 3 involves a binary classification task. The training dataset provided was in English but all subtasks mandated the evaluation of our model's zero-shot performance in three surprise languages: Bulgarian, North Macedonian, and Arabic and another fourth dataset in English. The goal during the testing phase was to explore our model's ability to generalize to these languages without explicit training.

This paper describes our participation to subtask 1, focusing on the detection of 20 persuasion techniques structured hierarchically within the textual content of memes. Inspired by successful approaches in multilabel text classification (Jurkiewicz et al., 2020; Tian et al., 2021), our strategy involved fine-tuning three language models i.e, BERT [`bert-base-uncased`], XLM-RoBERTa [`xlm-roberta-base`], and mBERT [`bert-base-multilingual-uncased`], followed by ensemble modeling using the mean aggregation technique using the English training set. To enhance performance, we used data augmentation through paraphrasing and adjusted the classification thresholds for each persuasion technique based on class-wise metrics optimised using the validation set using grid search. During testing, a zero-shot approach was implemented by translating the surprise language data into English.

At the shared task, our system demonstrated significant performance advantages over the baseline in all languages except Arabic, where the performance difference was not statistically significant. Our system's effectiveness, particularly in non-Arabic languages, underscores its potential for analyzing memes within disinformation campaigns, emphasizing the need for language-specific considerations in model development.

Section 2 provides an overview of the data utilized and offers insights into relevant prior research. Section 3 presents an overview of our classification pipeline, while Section 4 describes the experiments and data augmentation techniques that guided our final model decisions. Finally, Section 5 analyses the results of our model. All of the code used in the implementation of the models described in this paper is made available on GitHub.[1]

[1] https://github.com/CLaC-Lab/SemEval-2024-Task-4

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 175–180

2 Background

SemEval 2024 Task 4 (Multilingual Detection Of Persuasion Techniques In Memes) proposed 3 subtasks, out of which we participated in the first one. The goal of subtask 1 was to categorize the textual content of memes into one or several persuasion techniques. An inventory of 20 techniques was provided (eg: *Smears*, *Loaded Language*, *Slogans*) and were structured hierarchically, rendering the task a hierarchical multi-label classification problem.

2.1 Datasets

The SemEval organizers collected memes in English, Bulgarian, North Macedonian, and Arabic from their personal Facebook accounts, scraping public groups discussing politics, vaccines, COVID-19, gender equality, and the Russo-Ukrainian War. For subtask 1, the input data comprised the text extracted from these memes. The training (7k samples), validation (500 samples) and development (1k samples) sets included only English texts; whereas the test set was multilingual with 1500 samples for English, 426 samples for Bulgarian, 259 samples for North Macedonian and 100 samples for Arabic. All datasets were the provided in the form of JSON files. The orange bars in Figure 1 shows the distribution of the data for each persuasion technique in the training set. As Figure 1 shows some techniques, such as *Loaded Language* and *Smears*, had a substantial number of samples, while others like *Straw Man* and *Red Herring* were severely underrepresented.

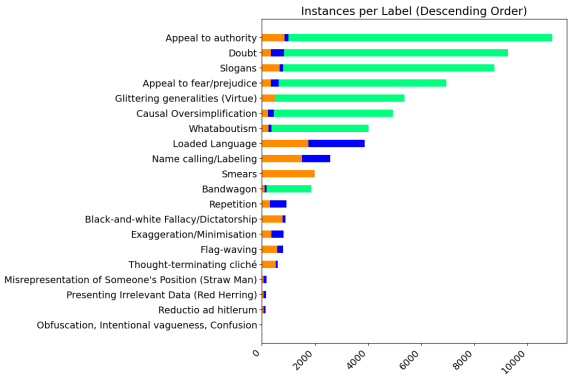

Figure 1: Distribution of the data for each persuasion technique in the SemEval 2024 (in orange), the `Comb-14k` (in orange + blue) and the `Para-54k` (in orange + blue + green) training datasets.

2.2 Previous Work

In the context of the SemEval 2020 Task 11 (Da San Martino et al., 2020), two subtasks were introduced addressing span identification of propagandistic textual fragments and a multi-label technique classification (TC) of propagandistic fragments using a corpus of ≈7k instances from the news domain. The subsequent SemEval 2021 Task 6 (Dimitrov et al., 2021) focused on the identification of propagandistic techniques from multimodal data including text and images from memes. This year's shared task build upon the 2021 task but included hierarchical metrics as well as a multilingual setting. The top-performing teams in 2020 and 2021, AppliCaAI (Jurkiewicz et al., 2020) and MinD (Tian et al., 2021) respectively, leveraged pre-trained language models and ensemble techniques to achieve top scores at the shared tasks. Inspired by these works, our methodology is also based on an ensemble of pre-trained language models.

3 System Overview

The aim of subtask 1 is to identify 0 or n persuasion techniques for each textual instance. Despite the hierarchical organization of the persuasion techniques, we opted to predicting solely the technique names (leaf nodes) and not their ancestor nodes. Figure 2 shows an overview of the classification pipeline we employed for this subtask. As shown in Figure 2, our methodology is based on fine-tuning three distinct pre-trained language models: BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and mBERT (Devlin et al., 2019). This fine-tuning process is conducted on augmented datasets.

3.1 Data Augmentation

As Figure 1 shows, some persuasion techniques have very few samples (eg: *Red Herring*, *Straw Man* only have 59 and 62 instances respectively) in the SemEval 2024 dataset (in orange). To mitigate the lack of data we took advantage of data augmentation strategies: The Technique Classification subtask from SemEval 2020 task 11 (Da San Martino et al., 2020) (See Section 3.1.1) and automatically generated paraphrases (See Section 3.1.2).

3.1.1 SemEval 2020 Data (`Comb-14k` dataset)

The Technique Classification (TC) subtask from the SemEval 2020 Task 11 (Da San Martino et al.,

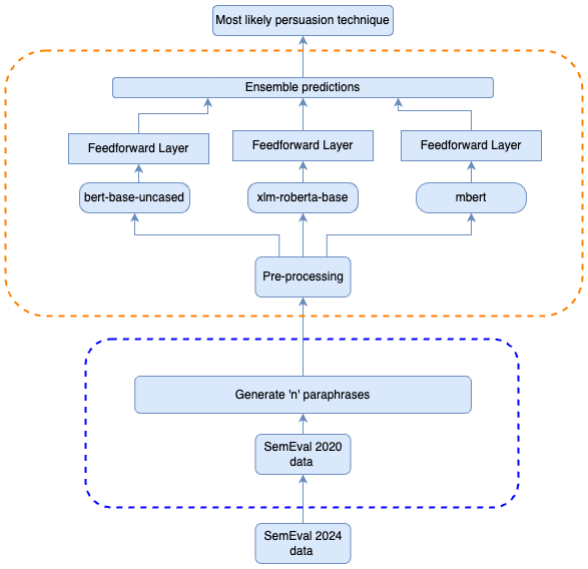

Figure 2: Schematic overview of our classification pipeline for the detection of persuasion techniques in memes.

2020) provided a dataset with ≈7k instances annotated with the same guidelines as this year's. In contrast to the 2020 task, this year's challenge featured a revised set of techniques compared to the 2020 inventory. In the 2020 TC dataset, a few techniques were merged into a single category due to lack of data, resulting in a list of 14 techniques. In the current year, an expanded inventory of 20 techniques was employed. To ensure consistency between the two sets, we preprocessed the 2020 TC dataset by splitting techniques that had previously been merged. For example, we singled out *Bandwagon* and *Reductio ad Hitlerum*, which had been merged into a single technique in the SemEval 2020 TC dataset.

We combined both datasets and fine-tuned models on this combined dataset. For easy reference in the rest of the paper, we call the combined dataset `Comb-14k`. Figure 1 (orange + blue) shows the resulting distribution of the persuasion techniques in this dataset.

3.1.2 Paraphrasing (`Para-28k`, `Para-52k` and `Para-54k` datasets)

Despite having almost doubled each class with the use of the 2020 TC dataset, some classes were still severely underrepresented; see Figure 1 (orange + blue). To address this, we augmented the dataset further by generating paraphrases for each instance. To generate paraphrases, we leveraged ChatGPT-3.5 turbo, setting the temperature to 0.7. This value aimed to introduce diversity in the paraphrases while maintaining relevance to the original instances.

For each instance in `Comb-14k`, we generated n paraphrases, then labeled these paraphrases with the same set of labels as the original instance. We experimented with n=1 and n=3. We call the resulting datasets `Para-28k` and `Para-52k`. The overall hierarchical F-score with the validation set given showed an increase when training with these datasets and $n = 3$ seemed to perform better than $n = 1$. A per-class analysis showed that not all classes benefited from the increase in support. For example, the persuasion technique *Bandwagon* increased its F1 from 0.17 to 0.29; whereas *Repetition* decreased its F1 from 0.56 to 0.31. We therefore identified the classes with improvement in F-score greater than 0.03 when using the `Para-52k` dataset compared to the `Comb-14k` dataset. These 8 techniques along with their increase in F-scores are shown in Table 1. This set of 8 techniques, referred to as benefited classes $\mathbf{B}$, formed the basis for our subsequent strategy. Since only these techniques seemed to benefit from the use of paraphrases, we only increased the number of paraphrases for these. Specifically, for all data instances d_i in `Comb-14k` labeled with techniques $\mathbf{T} = \{T_1, T_2, \ldots T_n\}$, for each $T_i \in \mathbf{B}$, we generate 10 paraphrases of d_i and label them with all techniques from $\mathbf{T} \cap \mathbf{B}$. This newly created dataset contained ≈54k instances, hence we call it `Para-54k`.

Figure 1 shows the distribution of instances for each technique in the `Para-54k` dataset (orange + blue + green), in comparison with the SemEval 2024 dataset and the `Comb-14k` dataset. As the figure shows, all datasets are severely imbalanced; something that we tried to address with the use of per-class custom thresholds (see Section 3.2).

3.2 Multi-label Classification

After creating the datasets, we preprocessed them using standard tokenization, then proceeded to fine-tune three distinct models: `bert-base-uncased`, `xlm-roberta-base`, and `bert-base-multilingual-uncased` in addition to an ensemble model, generated by averaging the predictions from all three models.

Additionally, we implemented thresholding in order to determine which techniques have a high enough score to be part of the output label set. We experimented with custom values for each of the techniques in order to address the data imbalance

Technique	Comb-14k		Para-52k		
	Support	F1	Support	F1	Δ F1
Bandwagon	169	0.17	676	0.29	0.12
Causal Oversimplification	449	0.00	1796	0.09	0.09
Appeal to fear/prejudice	631	0.26	2524	0.34	0.08
Doubt	843	0.08	3372	0.15	0.07
Appeal to authority	994	0.69	3976	0.74	0.05
Glittering generalities (Virtue)	488	0.38	1952	0.43	0.05
Slogans	796	0.42	3184	0.46	0.04
Whataboutism	366	0.32	1464	0.36	0.04

Table 1: Techniques that showed an improvement in F1 score when using n=3 paraphrases (i.e. Para-52k).

issue. We experimented with values ranging from 0.01 to 0.7 and picked the optimal values for each class based on the validation set (500 samples). These thresholds were applied to the scores obtained after passing the logits of each class through a sigmoid function. Table 2 shows the results of the validation with the optimal threshold for each class using the official scorer, which uses hierarchical metrics. As Table 2 shows, the best model with the validation set was the ensemble trained on the Para-52k dataset which reached an hierarchical F1 of 0.56. However, the ensemble model when trained on the Para-54k dataset, performed worse (hierarchical F1 of 0.54 with the validation set) than the ones that used lesser number of paraphrases (Para-28k and Para-52k). The ensemble, leveraging the collective insights of the three models, trained on the Para-52k emerged as the most effective in enhancing the overall system performance. Based on our results in the official leaderboard with the development set and validation results shown in Table 2, we chose to submit the ensemble model trained on the Para-52k dataset as it gave the best results with both the validation and the development set.

During the testing phase, the datasets in Bulgarian, North Macedonian, and Arabic were automatically translated to English for our model's zero-shot predictions. This was inspired by the approach of (Costa et al., 2023). The English test data was used as given.

4 Experimental Setup

4.1 Data Split and Augmentation

The training data provided in English initially comprised 7k samples. After combining it with 2020 TC dataset, the total increased to approximately 14k samples (Comb-14k). Subsequently, through paraphrase generation, the training dataset expanded to around 28k (Para-28k) when only 1 paraphrase per instance was used (n=1) and 52k (Para-52k), when n=3. Finally, the dataset with ten paraphrases for the benefited classes *B* reached approximately 54k samples (Para-54k). The original 500-sample validation set was used consistently for all our experiments. For the final submission, the ensemble model was trained on the union of (Para-52k) and the development set (1k samples), for a total of 53k samples.

4.2 System Pipeline and Training Details

The system pipeline code was implemented in PyTorch. The pre-trained models BERT [bert-base-uncased][2], XLM-RoBERTa [xlm-roberta-base][3], and mBERT [bert-base-multilingual-uncased][4] and their tokenizers were sourced from Hugging Face. Standard preprocessing, involving tokenization based on each model's tokenizer, was applied. Across all phases, models were trained for 10 epochs using the Adam optimizer with a learning rate of 2e-5. Batch sizes varied with BERT utilizing 128, and XLM-RoBERTa and mBERT using 64. A final feedforward layer with 20 logits (equal to the number of considered techniques) was added to each model. The Binary Cross Entropy with logits served as the loss function, with one-hot encoding applied to the true labels. For prediction, a sigmoid activation function was used on the logits, followed by thresholding. The ensemble model used an unweighted average of all predictions from the three individual models.

[2] huggingface.co/bert-base-uncased
[3] huggingface.co/FacebookAI/xlm-roberta-base
[4] huggingface.co/bert-base-multilingual-uncased

Training Set Used	Models	Validation Set	Development Set
Comb-14k	BERT	0.52	0.55
	XLM-RoBERTa	0.53	0.54
	mBERT	0.53	0.54
	Ensemble Model	0.53	0.56
Para-28k	BERT	0.55	0.57
	XLM-RoBERTa	0.57	0.54
	mBERT	0.50	0.53
	Ensemble Model	0.55	0.56
Para-52k	BERT	0.54	0.55
	XLM-RoBERTa	0.54	0.54
	mBERT	0.54	0.55
	Ensemble Model	**0.56**	**0.57**
Para-54k	BERT	0.48	0.51
	XLM-RoBERTa	0.54	0.55
	mBERT	0.51	0.53
	Ensemble Model	0.54	0.55

Table 2: Hierarchical F1 scores of our models, when trained on different English-language datasets for both the validation and development sets.

Language	Baseline	Our Score	Best Score
English	0.36865	0.57827	0.75427
Bulgarian	0.28377	0.44917	0.56833
North Macedonian	0.30692	0.39471	0.51244
Arabic	0.35897	0.38070	0.47593

Table 3: Comparison of the final hierarchical F1 scores obtained by our classification system, the best corresponding classification system in the shared task and the baseline in each given language.

ChatGPT-3.5 turbo[5] API with a temperature set to 0.7 was used for paraphrase generation. During testing, external languages were translated into English using the deep-translator API[6].
Throughout all phases hierarchical metrics were employed for task evaluation using the official scorer. On the other hand, standard precision, recall, and F-score metrics were used to assess the per class performance.

5 Results

The official performance results of our system are shown in Table 3, along with the baseline score and the score obtained by the best performing system on each language. As Table 3 shows, although our ensemble model was not among the top models, it reached significantly better performance than the baseline in all languages except Arabic, where the improvement was not significant. Overall, we stood at 22nd out of 33 participants for English, 12th out of 20 for Bulgarian, 11th out of 20 for North Macedonian and 11th out of 17 for Arabic.

6 Conclusion

This paper described the methodology used in our participation to the Semeval 2024 Task 4 subtask 1, focusing on hierarchical multi-label detection of persuasion techniques in meme texts. We used an ensemble model with three fine-tuned language models and incorporated additional strategies such as data augmentation through paraphrasing and classification thresholds fine-tuning based on class-wise metrics. During testing, our system significantly outperformed the baseline in all languages except Arabic, where the increase in performance was not significant. Analysis shows that the data augmentation and threshold fine-tuning may have introduced noise and exacerbating dataset imbalance.

[5] https://platform.openai.com/docs/models/gpt-3-5-turbo
[6] https://pypi.org/project/deep-translator/

Acknowledgements

The authors would like to thank the organisers of the SemEval shared task and the anonymous reviewers for their comments on the previous version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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.

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.

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.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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.

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

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1082–1087, Online. Association for Computational Linguistics.