Analyzing Persuasive Strategies in Meme Texts: A Fusion of Language Models with Paraphrase Enrichment

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

Abstract. This paper describes our approach to hierarchical multi-label detection of persuasion techniques in meme texts. Our model, developed as a part of the recent SemEval task, is based on fine-tuning individual language models (BERT, XLM-RoBERTa, and mBERT) and leveraging a mean-based ensemble model in addition to dataset augmentation through paraphrase generation from ChatGPT. The scope of the study encompasses enhancing model performance through innovative training techniques and data augmentation strategies. The problem addressed is the effective identification and classification of multiple persuasive techniques in meme texts, a task complicated by the diversity and complexity of such content. The objective of the paper is to improve detection accuracy by refining model training methods and examining the impact of balanced versus unbalanced training datasets. Novelty in the results and discussion lies in the finding that training with paraphrases enhances model performance, yet a balanced training set proves more advantageous than a larger unbalanced one. Additionally, the analysis reveals the potential pitfalls of indiscriminate incorporation of paraphrases from diverse distributions, which can introduce substantial noise. Results with the SemEval 2024 data confirm these insights, demonstrating improved model efficacy with the proposed methods.

Keywords: Language Models, Multi-label Classification, Persuasion Techniques.

1 Introduction

The recent SemEval-2024 shared task 4 [1] 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 and mandated the detection of 20 persuasion techniques structured hierarchically within the textual content of memes. On the other hand, subtasks 2 and 3 involved the analysis of multimodal memes that considered both textual and visual elements.

This paper describes the approach we used for our participation to subtask 1 and further analysis with the dataset after the shared task. The task provided a training dataset in English, but in addition to being tested on English, the task also

mandated the evaluation of our model's zero-shot performance in three surprise languages. The goal during the testing phase was not only to explore our model's ability to perform well in English but also generalize to other languages without explicit training.

Inspired by successful approaches in multilabel text classification [2, 3], our strategy involved fine-tuning three language models, i.e, BERT, XLM-RoBERTa, and mBERT, followed by ensemble modeling using mean aggregation. 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. During testing, a zero-shot approach based on machine translation was implemented to classify instances in the surprise languages. The official results of our system [1] demonstrated good performance advantages over the baseline in all languages except Arabic, where the increase in performance was not significant. Results also show that paraphrasing techniques enhances model performance but using a balanced training set is more beneficial than a larger unbalanced one. Furthermore, analysis highlights the potential downside of indiscriminately incorporating paraphrases from diverse distributions, as this can introduce notable noise into the system.

This paper is organized as follows: Section 2 reviews relevant previous work in the field of multi-label classification. Section 3 provides an overview of the task and the data utilized. Section 4 presents an overview of our classification pipeline along with the techniques used for data augmentation; while Section 5 describes experimental details. Finally, the analysis of our model's results is presented in Section 6; while Section 7 provides conclusions and outlines future work.

2   Previous Work

Several studies have explored multi-label classification of textual content. In 2019, [4] showed that Bi-GRUs with label-wise attention led to good performance, and the inclusion of domain-specific Word2vec and context-sensitive ELMo embeddings further boosted the performance on the EURLEX57K dataset that contained 57k English EU legislative documents. [5] introduced five innovative contrastive losses for multi-label text classification using the dataset from the SemEval 2018 Multi-label Emotion Classification (MEC) task [6] in English, Arabic, and Spanish that contained 8640 instances. All five contrastive learning methods notably enhanced the performance of the previous top-performing model, SpanEmO [7], for the MEC task. Among these approaches, the Jaccard Similarity Probability Contrastive Loss demonstrated the highest effectiveness on the English dataset, achieving F_{Macro} and F_{Micro} scores of 57.68 and 71.01, respectively.

In hierarchical multi-label classification (HMC), samples are assigned to one or more class labels within a structured hierarchy. Approaches to HMC can be divided

into local and global methods. Local methods use multiple classifiers, often overlooking the overall structure of the hierarchy. For example, [8] trained a multi-layer perceptron incrementally for each hierarchy level, using predictions from one level as inputs for the next. In contrast, global methods employ a single model to address all classes and implement various strategies to capture the hierarchical relationships between labels. One such approach [9] modeled the hierarchy as a directed graph and introduced hierarchy-aware structure encoders, using a bidirectional TreeLSTM and a hierarchy-GCN to extract and aggregate label structural information in an end-to-end fashion. [10] redefined hierarchical text classification (HTC) as a sequence generation task and developed a sequence-to-tree (Seq2Tree) framework to model the hierarchical label structure. Additionally, they created a constrained decoding strategy with a dynamic vocabulary to ensure label consistency.

In the context of the SemEval 2020 Task 11 [11], two subtasks were introduced addressing span identification of propagandistic textual fragments and a multilabel technique classification (TC) of propagandistic fragments using a corpus of 7k instances from the news domain. The subsequent SemEval 2021 Task 6 [12] focused on the identification of propagandistic techniques from multimodal data including text and images from memes. The top-performing teams in 2020 and 2021, ApplicaAI [2] and MinD [3] respectively, leveraged pre-trained language models and ensemble techniques to achieve top scores at the shared tasks. This year's shared task was built upon the 2021 task but included hierarchical metrics as well as a multilingual setting and a new training dataset of 7k instances in English. Inspired by the top models at [11] and [12], our methodology is also based on an ensemble of pre-trained language models but leverages paraphrase generation to further improve performance.

3   Labelling Persuasion Techniques

The goal of our model was to categorize the textual content of memes into one or several persuasion techniques. For example, given the training instance shown in Figure 1, the model needs to learn that the text *Don't expect a broken government to fix itself* should be labelled with the three techniques provided in the labels field.

Persuasion Techniques: The SemEval organizers provided an inventory of 20 persuasion techniques to be used as labels (eg: *Loaded Language*, *Slogans*, *Name calling/Labelling*) and were structured hierarchically as shown in Figure 2. This rendered the task a hierarchical multi-label classification problem and was therefore evaluated using hierarchical precision, recall and F measures.

Datasets: The dataset provided contained memes collected from online public groups discussing a variety of topics such as politics, vaccines, COVID-19, gender equality, and the Russo-Ukrainian War. For our task, only the text extracted

```
{
    "id": "79369",
    "text": "Don't expect a broken government to fix itself.",
    "labels": [
        "Loaded Language",
        "Slogans",
        "Name calling/Labeling"
    ],
    "link": "null"
}
```

Fig. 1. A sample training instance. The text is labelled with three techniques, *Loaded Language*, *Slogans* and *Name calling/Labelling*.

from these memes was used. The training (7k samples), validation (500 samples) and development (1k samples) sets included only English texts.

Figure 3 shows the distribution of instances for each persuasion technique in the training set. As the figure shows, some techniques, such as *Loaded Language* and *Smears*, had large number of instances in the training set (1750 and 1990 respectively); while others like *Straw Man* and *Red Herring* were severely underrepresented (62 and 59 instances respectively).

Figure 4 shows the distribution of labels per instances. As the figure shows, most of the instances (47%) were labelled with multiple techniques, 35% were labeled with only 1 technique and 18% had no labels at all.

Given the above English training set and hierarchical persuasion techniques, the goal of our model was to identify 0 or n techniques for each textual instance in English and in 3 surprise languages.

4   Proposed Approach

Figure 5 shows an overview of the classification pipeline we employed for this task. As shown in Figure 5, our methodology is based on fine-tuning three distinct pre-trained language models: BERT [13], XLM-RoBERTa [14], and mBERT [13] on augmented datasets.

4.1   Multi-label Classification

As Figure 5 shows, the data is first preprocessed using standard tokenization. Then we proceeded to fine-tune three distinct models: `bert-base-uncased`, `xlm-roberta-base`, and `bert-base-multilingual-uncased` which returned a probability distribution over the 20 techniques. These three model predictions were then pooled via averaging.

Despite the hierarchical organization of the persuasion techniques, we opted to predicting solely the technique names (leaf nodes in Figure 2) and not their ancestor nodes. However, to address the multi-label classification, we implemented
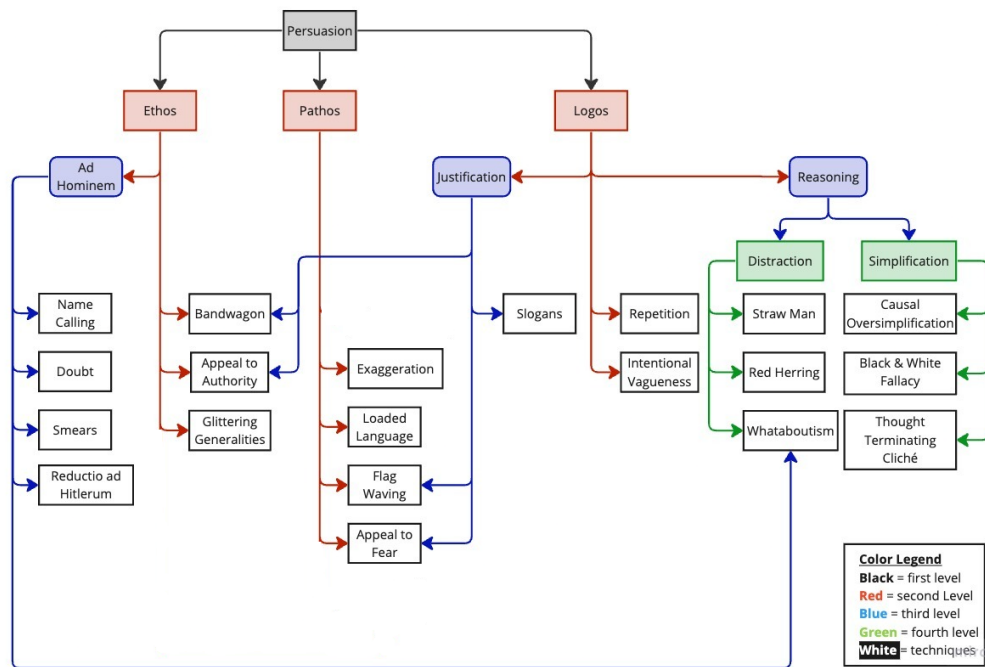
Fig. 2. Hierarchy of the persuasion techniques to be used to label the texts. Figure slightly modified from [1].

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 technique with values ranging from 0.01 to 0.7 and picked the optimal values for each class based on the validation set. These thresholds were applied to the scores obtained after passing the logits of each class through a sigmoid function.

To handle the three surprise languages, during the official testing phase system, the model trained only on English, would automatically translate the surprise language to English for our model's zero-shot predictions. This was inspired by the approach of [15].

4.2 Data Augmentation

To mitigate the lack of data we took advantage of two data augmentation strategies: an external dataset and automatically generated paraphrases.

External Dataset (`Comb-14k` dataset): The Technique Classification (TC) subtask from the SemEval 2020 Task 11 [11] provided a dataset of 7k instances from the news domain annotated with similar guidelines as this year's. In contrast to
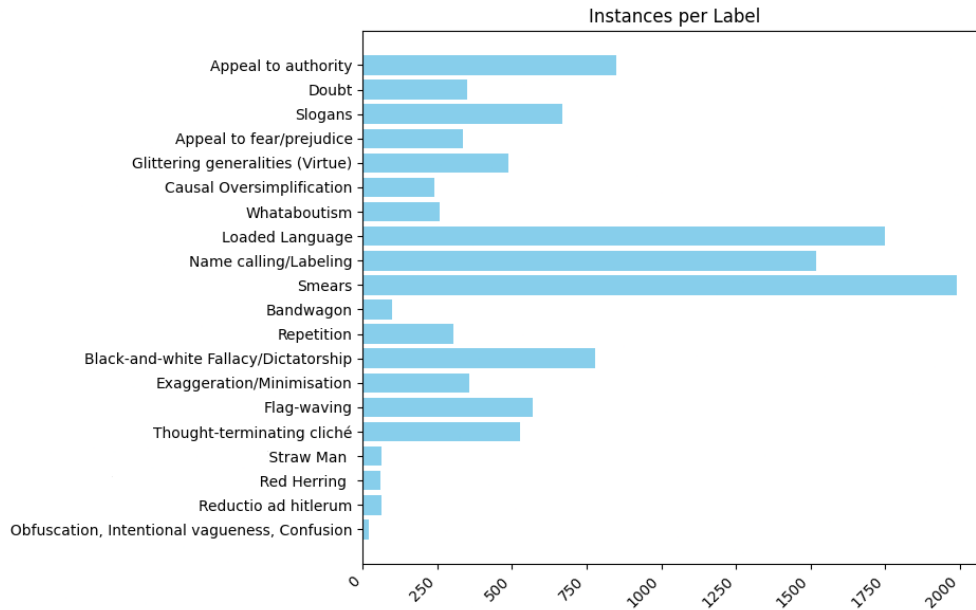
Fig. 3. Distribution of the data for each persuasion technique in the training set.

the 2020 task, this year's dataset covered a different domain and used a revised set of persuasion techniques compared to the 2020 inventory. Indeed, 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 pre-processed 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 considered two approaches to leverage the modified 2020 TC dataset. The initial option involved pre-training models on this dataset, followed by fine-tuning on the 2024 training data—an approach implemented by [3]. Another approach entailed combining both datasets and fine-tuning models on this combined dataset. We chose the latter method because the two datasets covered different genres and a joint training approach would likely enable the model to better adapt and grasp nuanced linguistic patterns across both. For easy reference in the rest of the paper, we call the combined dataset `Comb-14k`. Figure 7(a) (gray + pink) shows the resulting distribution of the persuasion techniques in this combined dataset.

Paraphrasing (`Para-*` datasets) Despite having almost doubled each class with the use of the 2020 TC dataset, some classes were still severely underrepresented; see Figure 7(a) (gray + pink). To address this, we took advantage of an LLM to gen-
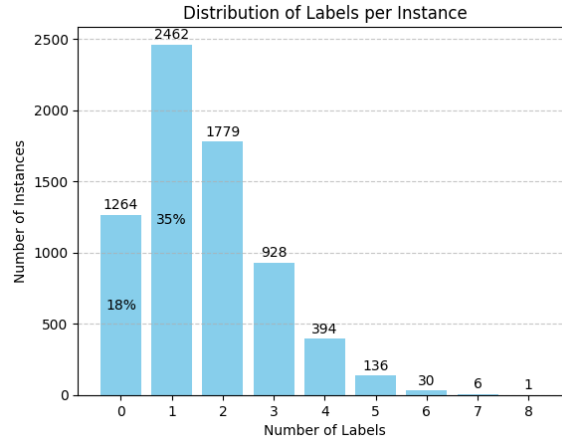
Fig. 4. Distribution of persuasion techniques per instance in the training set.

erate paraphrases for each training instance, then labeled these paraphrases with the same set of labels as the original instance. Our intuition was twofold. First, generating paraphrases would expose the model to a more extensive set of samples for each class, potentially improving its ability to discern subtle nuances within the data. Second, paraphrasing sentences could unveil hidden semantics, providing the model with a tool to identify propaganda techniques reliant on nuanced linguistic choices or phrasing. 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. Several datasets were created using this method:

Para-n1 and Para-n3: For each instance in `Comb-14k`, we generated n paraphrases. We experimented with n=1 and n=3 leading to datasets of 28k and 52k respectively, which we call `Para-n1` and `Para-n3` respectively.

The overall hierarchical F-score with the validation set given (500 instances) showed an increase when training with these datasets and $n = 3$ seemed to perform better than $n = 1$.

However, 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-n3` dataset compared to the `Comb-14k` dataset. These 8 techniques along with their increase in F-scores are shown in Figure 6. This set of techniques formed the basis for our subsequent strategy.

Para-Benef: Since only 8 techniques seemed to benefit from the use of paraphrases, we created a new augmented dataset by increasing the number of paraphrases
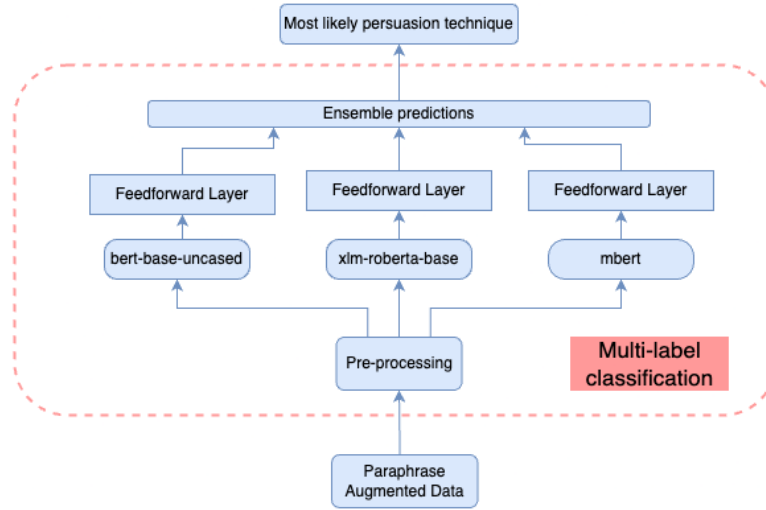
Fig. 5. Schematic overview of our classification pipeline for the detection of persuasion techniques in memes.

only for these techniques. Specifically, let **B** be the set of 8 techniques that benefited from paraphrases (see Figure 6), for all data instances d in `Comb-14k` labeled with techniques $\mathbf{T} = \{t_1, t_2, \ldots t_n\}$ (where $n \leq 20$), for each $t_i \in \mathbf{B}$, we generated 10 paraphrases of d and labeled them with all techniques from $\mathbf{T} \cap \mathbf{B}$. This newly created dataset called `Para-Benef`, contained 54k instances.

Figure 7(a) shows the distribution of instances for each technique in the `Para-Benef` dataset (gray + pink + violet), in comparison with the original training set and the `Comb-14k` dataset. As Figure 7 shows, all datasets are severely imbalanced. Our next dataset therefore tried to address this issue.

`Para-Bal:` Our last dataset used our paraphrase generation strategy to address the dataset imbalance. We rectified the underrepresented classes in the initial training dataset by augmenting them with paraphrases. The most frequent three techniques—*Smears*, *Name-calling/Labelling*, and *Loaded Language* had 1990, 1750, 1518 samples respectively. We thus aimed at reaching similar number of instances for the other techniques. We balanced the dataset by generating batches of 5 paraphrases for each other technique to reach around 1500 instances. This newly created dataset called `Para-Bal` contained 49k instances (see Figure 7(b)).

Table 1 shows the results of the validation with the optimal threshold for each class using the official SemEval scorer [1]. As Table 1 shows, the use of the original dataset (7k instances) achieved a hierarchical F1 of at most 0.49 on the development set; whereas all augmented training sets led to higher performances. The best model with both the validation and the development set was the ensemble trained
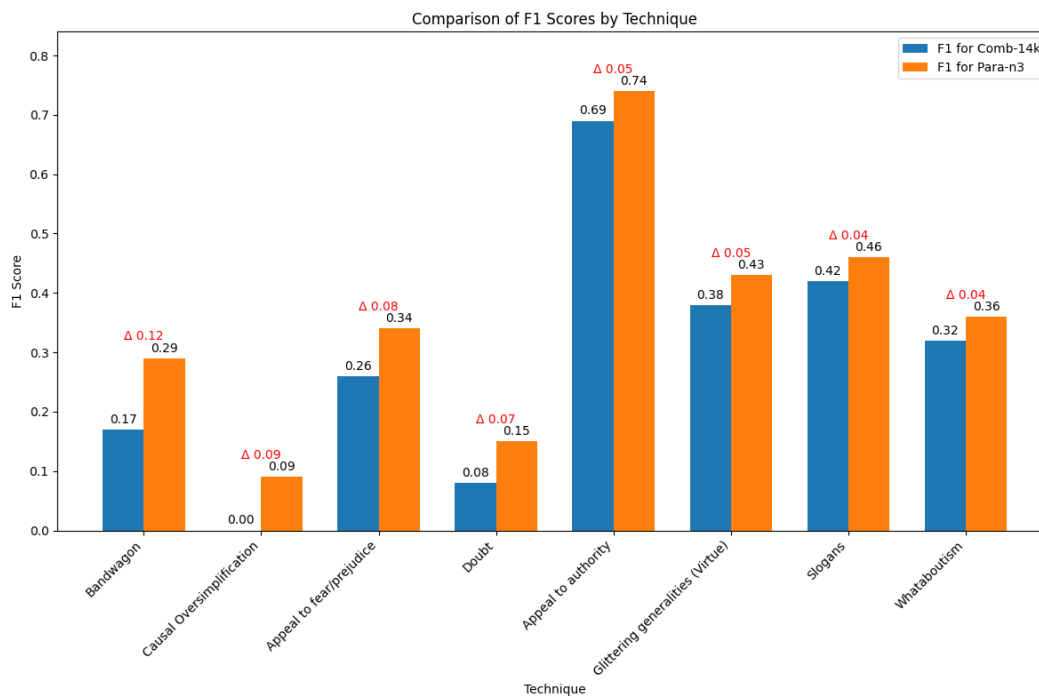
Fig. 6. Techniques that showed an improvement in F1 score with the validation set when using $n=3$ paraphrases (i.e. `Para-n3`) compared to `Comb-14k`.

on the `Para-Bal` dataset which reached an hierarchical F1 of 0.59 and 0.61 respectively. Surprisingly, `Para-Benef` which contained 10 paraphrases for the benefited techniques did not perform better than using only 3 paraphrases for all techniques (`Para-n3`). This suggests that the excessive inclusion of paraphrases from a different distribution (memes versus news) may have led to too much noise in the data.

5   Experimental Setup

5.1   System Pipeline and Training Details

The system pipeline code was implemented in PyTorch. The pre-trained models BERT [`bert-base-uncased`], XLM-RoBERTa [`xlm-roberta-base`], and mBERT [`bert-base-multilingual-uncased`] and their tokenizers were sourced from Hugging Face. All 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 persuasion techniques) was added to each model. The Binary Cross Entropy with logits served as the loss function, with one-hot encoding applied to
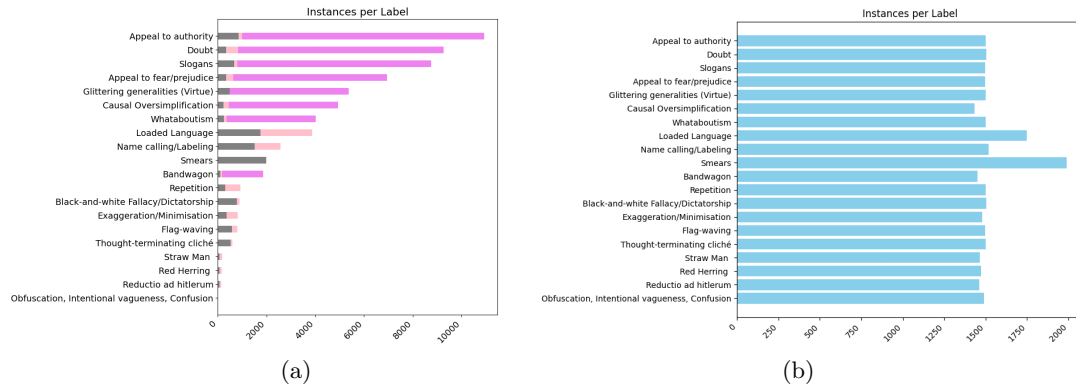
Fig. 7. Distribution of the data for each persuasion technique in (a) the original training set (in gray), the `Comb-14k` dataset (in gray + pink) and the `Para-Benef` dataset (in gray + pink + violet) (b) the `Para-Bal` dataset.

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. The ChatGPT-3.5 turbo[1] API with a temperature set to 0.7 was used for paraphrase generation. During testing, the surprise languages were translated into English using the deep-translator API[2].

6   Results and Analysis

For our official submission to SemEval, the `Para-Bal` dataset had not been created yet; hence our official results are based on the ensemble model trained on the union of `Para-n3` and the development set (1k samples), for a total of 53k samples. The three surprise languages were Bulgarian, North Macedonian and Arabic. The test set contained 1500 samples for English, 426 samples for Bulgarian, 259 samples for North Macedonian and 100 samples for Arabic. The official results of our system are shown in Figure 8, along with a baseline score that assigns the most frequent persuasion technique to all instances, and the score obtained by the best performing systems for each language [16, 17]. As Figure 8 shows, although our ensemble model did not reach the top performance for English (0.57 versus 0.75), it performed better than the baseline in all languages except Arabic, where the improvement was not significant.

Using the same testing protocol, we reproduced the results using the model trained with the balanced training dataset (`Para-Bal`). The results displayed in Figure 8 indicate an improvement in score with the English test set (0.62 versus 0.57). This again confirms the importance of a balanced dataset, and paraphrases

---

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

Table 1. Hierarchical F1 scores of our models, when trained on different English-language datasets for both the validation and development sets.

Training Set	Models	Validation Set	Development Set
Original (7k)	BERT	0.42	0.43
	XLM-RoBERTa	0.48	0.49
	mBERT	0.48	0.48
	Ensemble Model	0.45	0.46
Comb-14k (14k)	BERT	0.52	0.55
	XLM-RoBERTa	0.53	0.54
	mBERT	0.53	0.54
	Ensemble Model	0.53	0.56
Para-n1 (28k)	BERT	0.55	0.57
	XLM-RoBERTa	0.57	0.54
	mBERT	0.50	0.53
	Ensemble Model	0.55	0.56
Para-n3 (52k)	BERT	0.54	0.55
	XLM-RoBERTa	0.54	0.54
	mBERT	0.54	0.55
	Ensemble Model	0.56	0.57
Para-Benef (54k)	BERT	0.48	0.51
	XLM-RoBERTa	0.54	0.55
	mBERT	0.51	0.53
	Ensemble Model	0.54	0.55
Para-Bal (49k)	BERT	0.54	0.58
	XLM-RoBERTa	0.58	0.59
	mBERT	0.53	0.55
	Ensemble Model	**0.59**	**0.61**

based on the same distribution as the original texts. Indeed, although `Para-n3` is larger than `Para-Bal` (52k versus 49k), it is not balanced and contains paraphrases of instances from different genres (memes and news). However, surprisingly, the performance enhancement when using `Para-Bal` is not observed for the zero-shot classification of the surprise languages whose performance dropped significantly. For these languages, a larger training set, even with noisy out-of-distribution instances, leads to better results possibly due to the noise introduced by the automatic translation itself.

Compared to the other approaches at the 2024 edition of SemEval Task 4, the top performing team overall, OtterlyObsessedWithSemantics [17] designed a specialized classification head to enhance a Large Language Model. They organized the structure across several connected layers, enabling them to build upon earlier decisions in later, more detailed layers. They optimized the model's performance by systematically exploring different hyperparameters through grid-search. In addition, similarly to our approach, during the testing phase, they translated all the surprise language datasets into English. [16] developed a system using Chain-of-Thought based data augmentation methods, in-domain pre-training and ensemble strategy that combined the strengths of both RoBERTa and DeBERTa models.

Fig. 8. Comparison of the final hierarchical F1 scores obtained by our official model trained with `Para-n3`, the model trained with `Para-Bal`, the best corresponding models ([16] for English, [17] for Bulgarian, North Macedonian and Arabic) in the shared task and the baseline in each given language.

7   Conclusion and Future Work

This paper described our approach to hierarchical multi-label detection of persuasion techniques in meme texts. We used an ensemble model with three fine-tuned language models and incorporated data augmentation through paraphrasing from ChatGPT. We tested our approach through the SemEval 2024 Task 4 subtask 1 [1]. During testing, our system outperformed the baseline in all languages. Analysis of the results show the importance of dataset balancing and paraphrasing techniques in enhancing model performance. Despite having a smaller number of instances, the balanced dataset consistently outperforms its unbalanced counterparts, demonstrating the efficacy of balancing methods. Moreover, data augmentation improves model performance, as indicated by the under-performance of models trained on the original dataset (7k instances). Additionally, the results underscore the potential drawbacks of including paraphrases from diverse distributions, which may introduce significant noise into the system, potentially compromising overall effectiveness. This study prompts further inquiry into the specific drivers of performance improvement, whether it be dataset balancing or the inclusion of external data. Although our zero-shot approach exhibits limitations, it underscores the positive correlation between data volume and model performance, as illustrated by the supe-

rior performance of models trained on larger paraphrased datasets, such as `Para-n3` with 56k instances compared to `Para-Bal` with 49k instances

Moving forward, to boost performance, it would be interesting to measure the influence of the quality and similarity of the paraphrases on the performance. Moreover, incorporating hierarchical predictions, possibly at a second-level node, could improve scores further. Exploring the utilization of larger multilingual models alongside language-specific datasets and experimenting with various ensemble methods could be fruitful. Finally, considering the integration of adversarial training or self-supervised learning techniques might offer valuable avenues for improvement.

Acknowledgements

References

1. D. Dimitrov, F. Alam, M. Hasanain, A. Hasnat, F. Silvestri, P. Nakov, and G. Da San Martino, "Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes," in *Proceedings of the 18th International Workshop on Semantic Evaluation*, ser. SemEval 2024, Mexico City, Mexico, June 2024.
2. D. Jurkiewicz, Ł. Borchmann, I. Kosmala, and F. Graliński, "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*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1415–1424. [Online]. Available: https://aclanthology.org/2020.semeval-1.187
3. J. Tian, M. Gui, C. Li, M. Yan, and W. Xiao, "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)*, A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 1082–1087. [Online]. Available: https://aclanthology.org/2021.semeval-1.150
4. I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6314–6322. [Online]. Available: https://aclanthology.org/P19-1636
5. N. Lin, G. Qin, G. Wang, D. Zhou, and A. Yang, "An effective deployment of contrastive learning in multi-label text classification," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8730–8744. [Online]. Available: https://aclanthology.org/2023.findings-acl.556
6. S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 Task 1: Affect in tweets," in *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

7. H. Alhuzali and S. Ananiadou, "SpanEmo: Casting multi-label emotion classification as span-prediction," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 1573–1584. [Online]. Available: https://aclanthology.org/2021.eacl-main.135

8. R. Cerri, R. C. Barros, and A. C. de Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39–56, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022000013000718

9. J. Zhou, C. Ma, D. Long, G. Xu, N. Ding, H. Zhang, P. Xie, and G. Liu, "Hierarchy-aware global model for hierarchical text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1106–1117. [Online]. Available: https://aclanthology.org/2020.acl-main.104

10. C. Yu, Y. Shen, Y. Mao, and L. Cai, "Constrained sequence-to-tree generation for hierarchical text classification," *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247939715

11. G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1377–1414. [Online]. Available: https://aclanthology.org/2020.semeval-1.186

12. D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, and G. Da San Martino, "SemEval-2021 task 6: Detection of persuasion techniques in texts and images," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 70–98. [Online]. Available: https://aclanthology.org/2021.semeval-1.7

13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

14. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

15. N. F. Costa, B. Hamilton, and L. Kosseim, "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)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1613–1618. [Online]. Available: https://aclanthology.org/2023.semeval-1.223

16. D. Li, C. Wang, X. Zou, J. Wang, P. Chen, J. Wang, L. Yang, and H. Lin, "914isthebest at semeval-2024 task 4: Cot-based data augmentation strategy for persuasion techniques detection," in *Proceedings of the 18th International Workshop on Semantic Evaluation*, ser. SemEval 2024, Mexico City, Mexico, June 2024.

17. J. Wunderle, J. Schubert, A. Cacciatore, A. Zehe, J. Pfister, and A. Hotho, "Otterlyobsessed-withsemantics at semeval-2024 task 4: Developing a hierarchical multi-label classification head

for large language models," in *Proceedings of the 18th International Workshop on Semantic Evaluation*, ser. SemEval 2024, Mexico City, Mexico, June 2024.

Authors

Kota Shamanth Ramanath Nayak received B.Tech from Manipal Institute of Technology, India. Currently, he is pursuing his Master's in Computer Science from Concordia University, Canada. His research interests include Natural Language Processing, Discourse Analysis and Large Language Models.

Leila Kosseim obtained her PhD from the University of Montreal in 1995 on the topic of Natural Language Generation. Currently she is a professor in the Computer Science & Software Engineering (CSSE) Department at Concordia University in Montreal. Her research interests include Natural Language Processing, Artificial Intelligence and Text Mining.