NLPNCHU at SemEval-2024 Task 4: A Comparison of MDHC Strategy and In-domain Pre-training for Multilingual Detection of Persuasion Techniques in Memes

Shih-Wei Guo[1], **Yu-Ting Lin**[2], **Yu-An Lu**[3],**Yao-Chung Fan**[1*]
[1]Department of Computer Science and Engineering,
National Chung Hsing University, Taiwan
[2]Taipei Municipal Chenggong High School, Taiwan
[3]National Chupei Senior High School, Taiwan
{cometlcc,dong1214.mailbox,luyuam0}@gmail.com, yfan@nchu.edu.tw

Abstract

This study presents a systematic method for identifying 22 persuasive techniques used in multilingual memes. We explored various fine-tuning techniques and classification strategies, such as data augmentation, problem transformation, and hierarchical multi-label classification strategies. Identifying persuasive techniques in memes involves a multimodal task. We fine-tuned the XLM-RoBERTA-large-twitter language model[1], focusing on domain-specific language modeling, and integrated it with the CLIP visual model's embedding to consider image and text features simultaneously. In our experiments, we evaluated the effectiveness of our approach by using official validation data in English. Our system in the competition, achieving competitive rankings in Subtask1 and Subtask2b across four languages: English, Bulgarian, North Macedonian, and Arabic. Significantly, we achieved 2nd place ranking for Arabic language in Subtask 1.

1 Introduction

Propaganda and advertising serve as examples of persuasive discourse, which aims to change another's behavior, feelings, intentions, or views through communication, often in a one-sided manner (Lakoff, 1982). Hence, the context in which the communication occurs is crucial alongside the actual content being conveyed.

Memes, combining persuasive discourse on social media platforms, prove particularly effective. They spread ideas or emotions online and are a popular tool in misinformation campaigns, using various rhetorical and psychological techniques to influence users. Memes' visual components either reinforce or convey persuasive tactics, thus playing a significant role in shaping public opinion and attitudes. To address these challenges, SemEval-2024 introduced a shared task focusing on detecting persuasion techniques from multilingual memes (Dimitrov et al., 2024). This task defines a hierarchy directed acyclic graph (HDAG) to represent a meme's persuasive techniques and highlights the challenges and importance of understanding the nuances of digital persuasion.

This study proposes exploring the effectiveness of multi-dimensional hierarchical classification (MDHC) strategies in identifying persuasive techniques in memes, based on previous research and the successful application of MDHC strategies in real-world HDAGs. The results from a competition show our approach's effectiveness, ranking first in a specific subtask and competitively across others, demonstrating the potential of MDHC strategies in analyzing persuasive discourse.

2 Background

In this study, we explore the application of Hierarchical Multi-label Classification (HMC) in the context of persuasive techniques, by structuring them within a hierarchical multi-label framework. This approach allows for the simultaneous handling of both textual and visual data through multimodal modeling.

Recent research (Montenegro et al., 2023) has demonstrated the efficacy of the MDHC approach for HMC, noting its simplicity and ease of implementation. HMC, an advancement of Multi-label Classification (MC), is designed to predict multiple labels that are organized hierarchically from general to specific categories. The incorporation of hierarchical knowledge is found to significantly improve the performance of classifiers.

The MC, which is applicable in a wide range of areas, involves the challenge of predicting multiple interrelated category variables. As highlighted by Alfaro et al. (Alfaro et al., 2023) and Bielza and Larrañaga (bie, 2011), the complexity of MC compared to single-dimensional problems is primarily

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

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1868–1875
June 20-21, 2024 ©2024 Association for Computational Linguistics

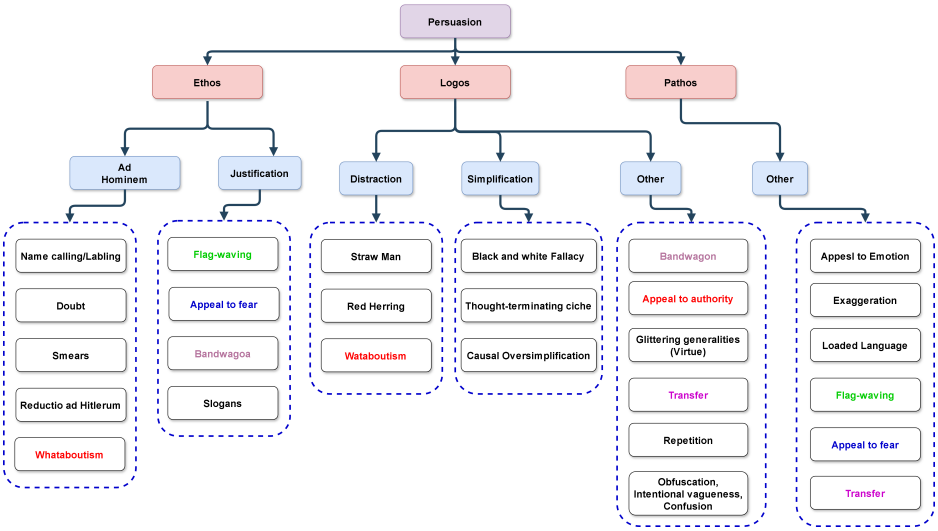

Figure 1: Hierarchy Multi-label Classification(HMC) with Persuasive Techniques

due to the vast combinations of class labels and the scarcity of relevant data.

We transform persuasive techniques into the HMC framework, this approach transforms persuasive technique graphs for application in specific subtasks. Given the necessity to analyze both textual and visual data for accurately identifying persuasive techniques, multimodal models become essential.

CLIP (Contrastive Language-Image Pre-Training) proposed by OpenAI (Radford et al., 2021), stands out for its independent text and image encoding capabilities, offering flexibility for various subtask types. In a study, (Kumar and Nandakumar, 2022) have suggested a range of techniques that combine textual and visual embedding vectors, leading to the effective detection of hateful memes. They also conducted various fusion experiments by switching different text encoders. Therefore, we refer to the authors' approach to combine the embedding vectors of the CLIP and the multilingual model with the aim of better adapting to Subtask 2ab, which involves tasks belonging to the multilingual domain and including datasets in three non-English languages (Bulgarian, North Macedonian, Arabic) in the test set.

3 Exploratory Data Analysis for Datasets

The dataset used in this study contains about 15,000 memes in English and other languages.

We have examined the label distribution for each task, and it is apparent that the data sets for subtask 1 and 2a are imbalanced, which differing numbers

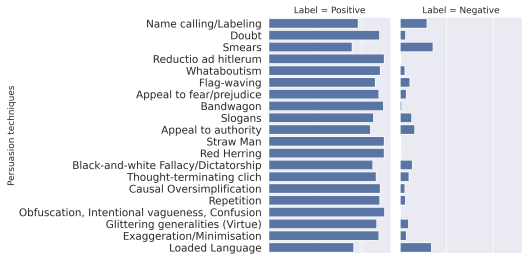

Figure 2: Data distribution of Persuasive Techniques on Subtask 1 Train Set

of Memes's persuasive techniques available for positive and negative are shown in Figure 2 and Figure 6. Further scrutiny, as delineated in Attachment Figure 5, it's evident that the datasets for Subtask 1 and 2a have a highly imbalanced distribution of data across 22 persuasion techniques, with Subtask 2a, in particular, showing a significant imbalance between positive and negative samples. We will describe how to address these imbalances in subsequent sections.

3.1 Transform the Structure of the Persuasive Techniques

The official release includes HDAG comprising 22 types of persuasive techniques. We have transformed this hierarchy into HMC. As shown in Figure 1, our reconstructed HMC has three levels: it includes 1 root node, the first layer has 3 child nodes, the second layer has 5 child nodes, and the bottom layer consists of 22 leaf nodes :

- **Root**: This describes whether a Meme image possesses any persuasive techniques.

- **First Layer Nodes**: There are 3 child nodes at this layer: Ethos, Logos, and Pathos. These nodes categorize the 22 types of persuasive techniques into 3 distinct classes of persuasive strategies.

- **Second Layer Nodes**: This layer includes 5 child nodes: Ad Hominem, Justification, Distraction, Simplification, and Other. We simplify the official hierarchy of persuasive techniques by using the "Other" node to encompass the Distraction and Simplification nodes, as they are redundant in the MDHC strategy.

- **Leaf Nodes**: There are 22 nodes at this level, corresponding to the 22 types of persuasive techniques that are the focus of this task. When a persuasive technique belongs to multiple categories in the first layer, it is represented by the same color.

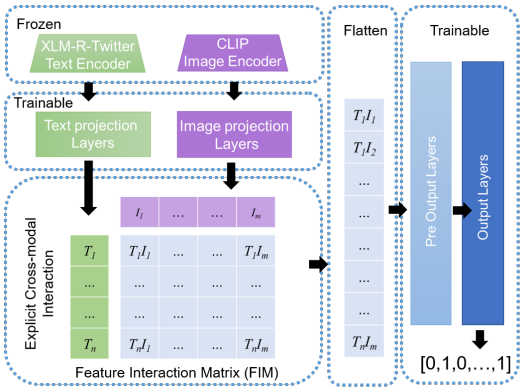

Figure 3: The Workflow for Multiclass Classification Task on the Multimodal Model

4 System Overview

In our research, we conducted an in-depth comparison of two MDHC strategies: Stacking+GC and Stacking+LCL, utilizing the same dataset for model training. The comparative analysis revealed that Stacking+GC demonstrated superior performance over Stacking+LCL. This superiority is attributed to its more effective handling of errors during the merging process of hierarchical levels, thereby enhancing the overall classification accuracy within the hierarchical structure of the data.

XLM-RoBERTA-large-twitter[1] For this system task, which is multilingual and specifically focused on the social media domain of Memes, we fine-tuned the domain-specific language model XLM-RoBERTA-large-twitter. This model was adjusted based on a corpus of over 156 million tweets in ten languages.

CLIP uses two distinct architectures as the backbone for encoding visual and textual datasets: image encoder, which represents the neural network architecture responsible for encoding images (e.g., ResNet or Vision Transformer), and text encoder, which represents the neural network architecture responsible for encoding textual information (e.g., BERT or Text Transformer). This structure is flexibly adapted to the subtasks of this project.

4.1 Detailed Description of the MDHC Strategy

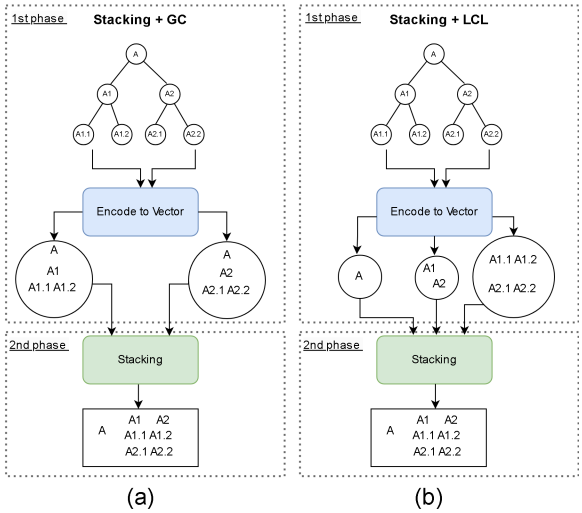

Figure 4: Illustration of the two DMC strategies. The feature vector is used on the 1^{st} phase as input, while the 2^{st} phase uses the outputs of the 1^{st} phase. it is to solve all the tasks associated with the internal nodes drawn inside the circle.

MDHC Strategies Incorporating the MDHC paradigm proposed by (Montenegro et al., 2023), our strategy selects an HMC classification approach suitable for this task, integrating both MC and HC strategies.

In MC strategy, we explore two solution algorithms: **Local Classifier per Level (LCL)**, which creates a model for each level of the hierarchy, and **Global Classifier (GC)**, a model that learns and predicts across the entire class hierarchy.

For HC strategy, we adopt *Stacking*, as introduced by (Wolpert, 1992), leveraging predictions from other labels to refine initial predictions, us-

ing average confidence scores to identify specific persuasive techniques with a threshold of 0.5 for determination.

To compare MDHC classification strategies, we integrate two MC strategies with an HC strategy, resulting in two distinct approaches.

Stacking + GC : this strategy applies the GC strategy to each dimension in the 1^{st} phase, utilizing feature vectors as inputs. The 2^{nd} phase employs the Stacking strategy, concatenating the probability vectors of the 1^{st} phase classifier's predictions for output. This strategy is shown in a Figure 4 (a)

Stacking + LCL : this strategy applies the LCL strategy to each dimension in the 1^{st} phase. In the 2^{nd} phase, it also adopts the Stacking strategy, using concatenated probability vectors from the first stage's predictions, with each circle representing a classifier addressing the classification problem of listed parent nodes. This strategy is shown in a Figure 4 (b)

4.2 Internal Negative Data Augmentation

In typical datasets, there is usually a similar ratio of positive to negative samples, even though they may not be evenly distributed. However, in the data distribution for Subtask 2b, there are only 2 samples that do not contain persuasive elements. Generally, one could use the PTC[2] dataset to augment this. However, the PTC dataset consists of news sentences with 18 types of persuasive techniques, and in hierarchical multi-label classification, the multiple labels have complex relationships.

When attempting to augment data by adding more examples for underrepresented labels, one must navigate the complex interplay between these labels carefully. Simply increasing the number of samples for a specific label can inadvertently exacerbate the imbalance for others. For example, if we augment the dataset with more instances of the 'Whataboutism' technique without considering its relationship with other techniques like 'Loaded Language' or 'Flag-waving,' we might skew the dataset further, making it even more challenging to train a balanced and accurate classifier.

To address this issue, we used two MDHC strategies. Firstly, we divided the entire hierarchy of persuasive techniques into six tasks based on the second-layer parent nodes: Ad Hominem, Justification, Distraction, Simplification, and Other (which was further split into two tasks). Each task independently applies the MC strategy, considering its specific persuasive techniques as positive examples and the others as negative examples, which approach we called **Internal Negative Data Augmentation(INDA)**, not only offers effective negative examples but also ensures consistency in labeling across various datasets. Ultimately, these six MC tasks determine the classification of the top-level parent node (indicating the presence of persuasive techniques) through a voting mechanism.

However, the INDA, while addressing the issue of imbalanced label distribution among samples without persuasive techniques, introduces the Long Tail Distribution problem. Long Tail Distribution is a probability distribution model characterized by lower probability density in its tail. In many cases, the distribution's right tail is considered more significant, but the left tail has a higher probability density. As shown in Attachment Figures 7, 8 and 9, the MDHC classification strategy we employed results in the positive distribution of a particular persuasive technique type being concentrated in the tail, forming a left-skewed long tail distribution.

The long-tail distribution presents two main challenges: Label co-occurrence and Dominance of negative labels:

Label co-occurrence : texts are often associated with multiple persuasive techniques simultaneously, making it difficult to accurately sample individual categories.

Dominance of negative labels : a text may only be associated with a small subset of persuasive techniques, resulting in the majority of labels being negative. However, Binary Cross-Entropy (BCE) treats positive and negative associated categories equally, leading to a shift in the boundary of negative associations.

To address the issues of Label co-occurrence and Dominance of negative labels, we introduce the Distribution-Balanced Loss (DBL) proposed by (Wu et al., 2020), the loss function employs re-balanced weighting and negative-tolerant regularization to mitigate the challenges posed by Label co-occurrence and Dominance of negative labels.

Therefore, without relying on external data augmentation, we utilize the MDHC strategy combined with the DBL loss function to address the sample imbalance issues in Subtask 1 and 2a. Additionally, we also introduce the DBL loss function to tackle the long-tail distribution problem.

[2]https://propaganda.math.unipd.it/ptc/

4.3 Meme with Multimodal Learning

In Subtask 2a and 2b, we employ the CLIP multimodal model, which includes an image encoder and a text encoder. The image encoder is responsible for encoding images, while the text encoder handles encoding textual information. At this stage, we utilize the XLM-RoBERTA-twitter fine-tuned on Subtask 1 as the text encoder because this model already possesses a certain understanding of meme text. For the image encoder, we use the pre-trained CLIP image encoder provided by the official source. Through Feature-wise Linear Modulation (FIM), these two encoders encode to obtain a representation embedding vector containing both image and text, enabling the model to effectively comprehend memes reliant on the relationship between text and image, such as persuasive techniques like Transfer and Appeal to (strong) Emotions.

Specifically, as depicted in Figure 3, Subtask 2ab involves encoding Meme images using the image encoder (I) and Meme text using the text encoder (T). Thus, we obtain sets of image encoding vectors $I_1...I_N$ and text encoding vectors $T_1...T_N$. We compute the FIM by multiplying these two vectors to obtain a new set of feature vectors. Subsequently, we attach a linear layer for classification at the end of the model to output the correct classifications. Specifically, in Subtask 2a, the learning objective is multi-label (MC) persuasion techniques classification, while in Subtask 2b, the learning objective is binary classification to identify whether it contains persuasive techniques or not.

5 Experiment and Evaluate

In our experiments, we employed three MDHC classification strategies on subtasks. All tasks utilize the HierarchyF_1 evaluation metric, which is the unified evaluation metric provided by the official source. In the experiment setting, refer to Appendix A for details of the relevant parameters.

In **Subtask 1**, we compared the performance of two language models, XLM-RoBERTA and XLM-RoBERTA-Twitter, across Subtask 1 and 2a, to assess the impact of domain-specific pre-training. We explored three classification strategies: GC, Stacking + GC, and Stacking + LCL, to identify the most effective approach for Subtask 1. In **Subtask 2a**, we evaluated two multimodal models by combining CLIP with XLM-RoBERTA-Twitter from Subtask 1. The goal was to improve the comprehension of Meme's persuasive techniques by utilizing a fine-tuned text encoder. Additionally, we employed GC, Stacking + GC, and Stacking + LCL strategies to determine which one is more effective for subtask 2a. Our primary focus was on improving classification performance in a multimodal setting. In **Subtask 2b**, we explored multimodal model and classification strategy using the CLIP + XLM-RoBERTA-Twitter combination from subtask 1, applied to a binary classification framework. Utilizing a balanced dataset of hate Memes collected by Meta AI, as referenced in the Harmful Memes Dataset (Kiela et al., 2020), the goal was to train the model on balanced samples to prevent bias effectively.

Table 1: Performance for Subtask 1 in Validation Set

Models	Strategy	Metrics		
		H-F1	H-Prec	H-Rec
Baseline	Official	0.3651	0.4573	0.3038
XLM-R Large	GC	0.5594	0.4635	0.6335
	Stacking + LCL	0.5995	0.5244	0.6909
	Stack + GC	0.6262	0.5907	0.6646
XLM-R-Large Twitter	GC	0.5580	0.6367	0.6503
	Stacking + LCL	0.6310	0.6062	0.6764
	Stacking + GC	**0.6689**	**0.6843**	**0.7451**

Table 2: Performance for Subtask 2a in Validation Set

Models	Strategy	Metrics		
		H-F1	H-Prec	H-Rec
Baseline	Official	0.4589	0.6820	0.3457
CLIP + XLM-R Large	GC	0.5214	0.6320	0.5775
	Stacking + LCL	0.6265	0.6343	0.5947
	Stack + GC	0.6675	0.7598	0.6107
CLIP + XLM-R-Large Twitter	GC	0.5581	0.6369	0.6103
	Stacking + LCL	0.6567	0.6459	0.6178
	Stacking + GC	**0.7134**	**0.7652**	**0.6418**

Table 3: Performance for Subtask 2b in Validation Set

Models	Strategy	Metrics	
		F1 macro	F1 micro
Baseline	Official	0.2500	0.3333
CLIP + XLM-R-Large	Binary Classification	0.7618	0.7947
CLIP + XLM-R-Large Twitter	Binary Classification	**0.8023**	**0.8216**

Results Our study's evaluation of the official validation set showcases the impactful of domain knowledge in enhancing model performance for persuasive technique identification across various subtasks. **Subtask 1**: As shown in Table 1, the model fine-tuned with domain knowledge performs significantly better in identifying persuasive techniques. Among the three classification strategies, Stack + GC demonstrates superior performance compared to the other two strategies. **Subtask 2a**:

The evaluation results for this task, as depicted in Table 2, align with Subtask 1. The model fine-tuned with domain knowledge outperforms others in identifying persuasive techniques. Among the three classification strategies, Stack + GC exhibits superior performance. **Subtask 2b**: In Table 3, this task involves binary classification. We utilized a multimodal model for binary classification and achieved competitive scores through external data augmentation methods.

Table 4: Performance for Subtask 1 in Test Set

Languages	Method	Metries		
		H-F1	H-Prec	H-Rec
English	Baseline	0.36865	0.47711	0.30036
	Ours	**0.66271**	**0.60990**	**0.72552**
Bulgarian	Baseline	0.28377	0.31881	0.25567
	Ours	**0.51744**	**0.53578**	**0.50031**
North Macedonian	Baseline	0.30692	0.31403	0.30012
	Ours	**0.46165**	**0.54622**	**0.39975**
Arabic	Baseline	0.35897	0.35000	0.36842
	Ours	**0.47500**	**0.42817**	**0.53333**

Table 5: Performance for Subtask 2a in Test Set

Languages	Method	Metries		
		H-F1	H-Prec	H-Rec
English	Baseline	0.44706	0.68778	0.33116
	Ours	**0.70677**	**0.78164**	**0.64498**
Bulgarian	Baseline	0.50000	**0.80428**	0.36276
	Ours	**0.54864**	0.70691	**0.44828**
North Macedonian	Baseline	**0.55525**	**0.90219**	**0.40103**
	Ours	0.48707	0.70575	0.37185
Arabic	Baseline	**0.48649**	**0.65000**	0.38870
	Ours	0.48323	0.59466	**0.40698**

Table 6: Performance for Subtask 2b in Test Set

Languages	Method	F1 macro	F1 micro
English	Baseline	0.25000	0.33333
	Ours	**0.78803**	**0.82167**
Bulgarian	Baseline	0.16667	0.20000
	Ours	**0.64706**	**0.82000**
North Macedonian	Baseline	0.09091	0.10000
	Ours	**0.52000**	**0.79000**
Arabic	Baseline	0.22705	0.29375
	Ours	**0.58518**	**0.59375**

In the test set of the competition, we propose a method that has demonstrated competitive performance on the official competition leaderboard. As shown in Table 4, our method outperforms all official baselines in Subtask 1, which involves the identification of meme persuasion techniques in four different languages at the text level. Our method ranks 4th on average across four languages (English, Bulgarian, North Macedonian, and Arabic), with rankings of 6th in English, 3rd in Bulgarian, 4th in North Macedonian, and 2nd in Arabic. These results indicating our method is competitive.

In the multimodal task, as shown in Table 5, We observe the performance of our method in Subtask 2a, where it demonstrates competitiveness in English memes. We attribute this to two main reasons. Firstly, our method only undergoes text-level domain pretraining on the English memes provided in the training dataset. As a result, it lacks the necessary representation capabilities for low-resource languages, such as North Macedonian. based on the above, our method does not improve cross-linguistic abilities; instead, it relies on the language representation capabilities of the multilingual model. Therefore, this makes it challenging to identify cross-lingual fine-grained persuasion techniques. Secondly, the inherent cultural differences in various languages may lead to discrepancies in the same set of memes presented in different languages, indicating a potential issue of cultural divergence.

Finally, in Subtask 2b, as shown in Table 6, although our proposed method falls short in the fine-grained cross-linguistic meme persuasion techniques identification, it remains competitive in tasks involving the identification of whether a meme, combining text and image, contains one of the persuasion techniques. Our method ranks 6th on average across four languages (English, Bulgarian, North Macedonian, and Arabic), with rankings of 7th in English, 5th in Bulgarian, 6th in North Macedonian, and 5th in Arabic. This demonstrates the competitive edge of our method in a multimodal context.

6 Conclusion

We conducted a detailed analysis of the HDAG containing persuasive techniques, transforming it into an HMC task. We also explored two MDHC strategies and highlighted the importance of addressing the long-tail distribution issue, proposing the use of the DBL loss function to mitigate this issue in HMC tasks. Regarding the models, we recommend utilizing domain-specific pre-training to detect memes containing persuasive elements and the effectiveness of domain-specific training was demonstrated across various experiments. Finally, we achieve competitive results in the competition.

References

2011. Multi-dimensional classification with bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727.

Juan C Alfaro, Juan A Aledo, and José A Gámez. 2023. Multi-dimensional bayesian network classifiers for partial label ranking. *International Journal of Approximate Reasoning*, page 108950.

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

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robin Tolmach Lakoff. 1982. Persuasive discourse and ordinary conversation, with examples from advertising. *Analyzing discourse: Text and talk*, pages 25–42.

C Montenegro, R Santana, and JA Lozano. 2023. Introducing multi-dimensional hierarchical classification: Characterization, solving strategies and performance measures. *Neurocomputing*, 533:141–160.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer.

A Implementation Details

During training, we use AdamW as the optimizer and an initial learning rate of 2e-5 for XLM-Roberta-twitter and 1e-4 for CLIP models. with a batch size of 32 and text max length set to 128 on subtask 1 and a batch size of 16, image size set to 224, and text max length set to 128 on subtask 2a and 2b. with all subtasks, the maximum number of epochs is set to 50. All experiments are conducted using two NVIDIA TITAN RTX GPUs

B Data Distribution Details

B.1 Data Distribution of the Persuasion Techniques

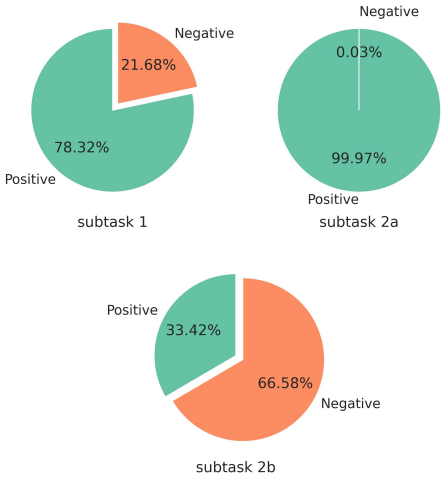

Figure 5: Ratio of the Persuasion Techniques on Subtask 1,2a and 2b

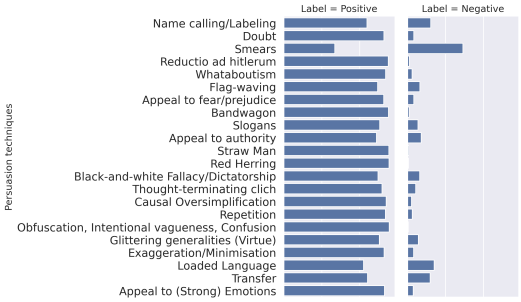

Figure 6: Data Distribution of Persuasion Techniques on Subtask 2a Train set

B.2 Data Distribution of MC for Persuasion Techniques in the MDHC strategy

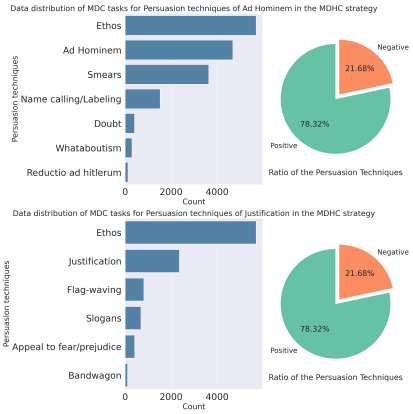

Figure 7: Data Distribution of MC for Persuasion Techniques of Ethos in the MDHC Strategy on Subtask 2a

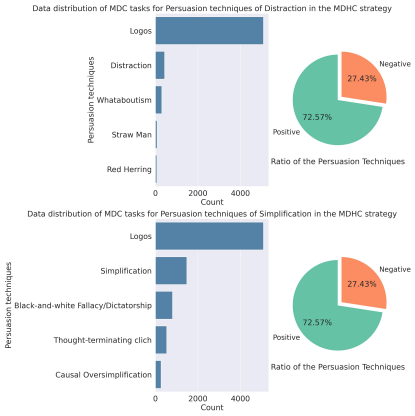

Figure 8: Data Distribution of MC for Persuasion Techniques of Logos in the MDHC Strategy on Subtask 2a

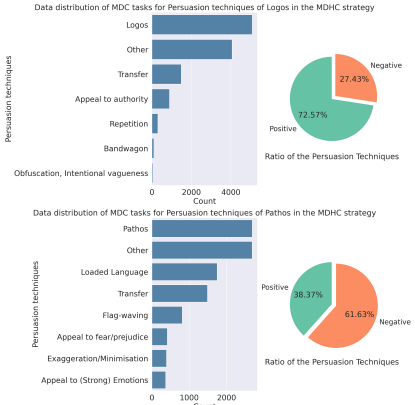

Figure 9: Data Distribution of MC for Persuasion Techniques of Pathos and Logos in the MDHC Strategy on Subtask 2a