Snarci at SemEval-2024 Task 4: Themis Model for Binary Classification of Memes

Luca Zedda and **Alessandra Perniciano** and **Andrea Loddo** and
Cecilia Di Ruberto and **Manuela Sanguinetti** and **Maurizio Atzori**
Department of Mathematics and Computer Science, University of Cagliari
Via Ospedale 72, Cagliari (Italy)
{luca.zedda, alessandra.pernician, andrea.loddo,
cecilia.dir, manuela.sanguinetti, atzori}@unica.it

Abstract

This paper introduces an approach developed for multimodal meme analysis, specifically targeting the identification of persuasion techniques embedded within memes. Our methodology integrates Large Language Models (LLMs) and contrastive learning image encoders to discern the presence of persuasive elements in memes across diverse platforms. By capitalizing on the contextual understanding facilitated by LLMs and the discriminative power of contrastive learning for image encoding, our framework provides a robust solution for detecting and classifying memes with persuasion techniques. The system was used in Task 4 of Semeval 2024, precisely for Substask 2b (binary classification of presence of persuasion techniques). It showed promising results overall, achieving a Macro-$F_1 = 0.7986$ on the English test data (i.e., the language the system was trained on) and Macro-$F_1 = 0.66777/0.47917/0.5554$, respectively, on the other three "surprise" languages proposed by the task organizers, i.e., Bulgarian, North Macedonian and Arabic. The paper provides an overview of the system, along with a discussion of the results obtained and its main limitations.

1 Introduction

In recent years, the natural language processing (NLP) community has witnessed an ever-growing number of contributions aimed at identifying and analyzing various forms of harmful language found on the Web, including offensive language (Zampieri et al., 2020), hate speech (Basile et al., 2019; Röttger et al., 2022)—also comprising misogyny and transphobia (Nozza et al., 2022; Kirk et al., 2023), and propaganda techniques (Da San Martino et al., 2019). These linguistic phenomena not only harm civil debate but can also fuel the polarization and radicalization of users' opinions.

In an increasingly multimodal context, particular attention has also been paid to memes (Dimitrov et al., 2021), which, due to their virality and communicative immediacy, can easily become key tools in online disinformation campaigns. Therefore, the development of techniques to effectively classify possible nuances of information manipulation within these forms of content sharing assumes a central role in online disinformation research.

This motivated our participation in the SemEval 2024 Task 4[1] (Dimitrov et al., 2024), which focuses on "Multilingual Detection of Persuasion Techniques in Memes". The task aims to develop models capable of identifying rhetorical and psychological techniques employed in memes to influence users' opinions. Our team participated in Substask 2b, consisting of a binary classification problem to determine whether a meme contains at least one persuasion technique among the predefined set of 22 techniques. The dataset released to participants is made up of memes with textual content in English. However, to assess the robustness of the systems during the evaluation phase, the organizers made test sets available in three other languages besides English, i.e., Arabic, Bulgarian, and North Macedonian.

For the purpose of this task, we developed a system that combines Large Language Models (LLMs) and contrastive learning image encoders to discern the presence of persuasive elements in memes across diverse platforms. The following sections will thus describe the system architecture and its deployment in the task. A discussion of the results obtained and of some most recurring errors will also be proposed, aiming to highlight possible research paths for the further improvement of the approach.

[1] https://propaganda.math.unipd.it/semeval2024task4/index.html

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 853–858

2 System Overview

The proposed method, named "Themis", is a modular neural network architecture designed to analyze multimodal data, specifically targeting memes that often contain both textual and visual elements.

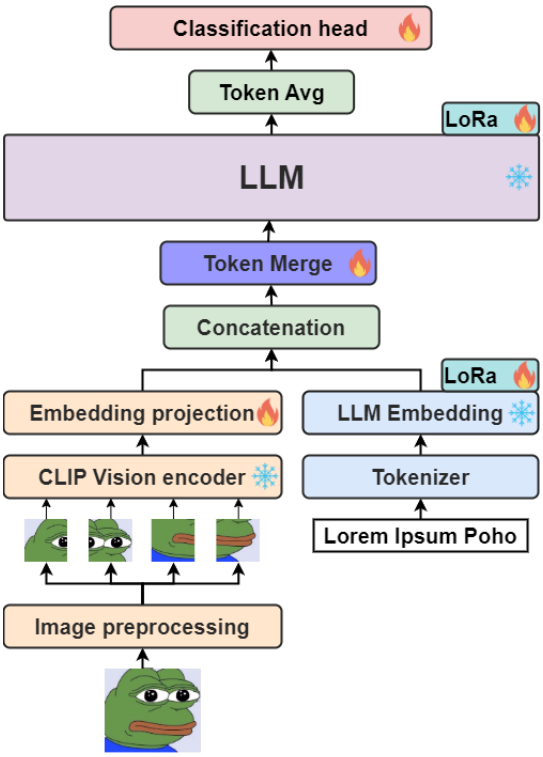

Figure 1: Themis model architecture.

2.1 Model Composition

Our Themis model comprises various interoperating components, as outlined in Figure 1. In this section, we shall examine their respective roles within the architecture.

Image Preprocessing. The meme image initially undergoes processing through the image processor of the Image Embedding Model, which standardizes and resizes it to conform to the model's specifications. Subsequently, the image is segmented into uniform patches of predetermined dimensions for subsequent processing. This preprocessing step guarantees that the image is adequately prepared for further analysis and feature extraction by the Image Embedding Model.

Image Embedding Model. Themis integrates an image embedding model for the extraction of features from meme images. This model is responsible for processing pixel values and extracting significant representations from the images.

Image Embedding Projection. An image embedding projection is applied to the features extracted from the image embedding model. This projection serves as a method to project image features in LLM-compatible size.

Large Language Model. Themis uses a language model to handle textual and visual inputs associated with memes. The LLM serves as the core of our model sequentially aligning tokens to an embedding space related to the persuasion detection task.

Token Merger Module. Themis employs a Token Merger module to merge tokens representing both image and textual features, enabling the model to attend to pertinent information within the images. This functionality allows the model to focus on salient aspects during meme processing. While drawing inspiration from the Patch Merger module (Renggli et al., 2022), our approach distinguishes itself by integrating both modalities. The Token merger learns a weight matrix that computes token scores based on representations and normalizes them using softmax. Subsequently, these weights are used to reduce the number of tokens through matrix multiplication. Ideally, this module aggregates similar tokens together, regardless of their original position. To address scale mismatches, layer normalization is applied post-merging, facilitating rapid adaptation through fine-tuning.

Token Average. We employ the token averaging technique, which involves extracting tokens from the LLM, to derive our final prediction. This strategy is designed to generate a single, semantically dense embedding, facilitating seamless processing by a classification head for obtaining the class prediction.

Classification Head. Themis incorporates a classification head to predict whether a meme contains specific persuasion techniques. This head takes the fused multimodal features and generates predictions based on the learned LLM representations.

2.2 Model Freezing and Low-Rank Adaptation (LoRA) Weights

The Themis model uses freezing techniques to control the training of certain parameters. Specifically, both the image embedding model and the language model are frozen during training. This ensures that

the pre-trained weights of these models are not updated, preserving the learned representations.

Additionally, Themis employs LoRA (Hu et al., 2022) weights to enhance its capabilities. LoRA weights are incorporated into the Image Embedding Projection layer and LLM model to introduce long-range interactions between tokens and patches, facilitating the capture of global context and improving overall performance in meme analysis tasks.

3 Experiment Setup

The experiments were executed on a workstation featuring an Intel Core i7-12700 @ 2.1GHz CPU, 32 GB RAM, and an NVIDIA RTX3060 GPU with 12GB of memory. Among the different experiments, one main issue is denoted by the limited availability of VRAM; this issue not only limited our approach to smaller LLM and Image encoders but also limited batch size. Our experiments aim to enable efficient prediction even in such low-end system requirements. As a result, we opted for a pre-trained Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) encoder as our main image encoder. Specifically, we used both CLIP Base[2] and CLIP Large[3] in our experiments. For the textual part, instead, we used TinyLlama[4] (Zhang et al., 2024), Phi-1.5[5] (Li et al., 2023) and Phi-2[6]. Table 1 depicts the full set of selected Image encoders and LLMs that suited our requirements.

Tipology	Model	# Params (B)
Image Encoder	CLIP Base 32	0.15
	CLIP Large 14	0.42
LLM	Phi-1.5	1.3
	Phi-2	2.7
	TinyLlama	1.1

Table 1: List of LLMs and Image Encoders used for our experiments.

For each combination of image and text models, the system was trained for 20 epochs, using a batch of 2 and a learning rate of $1e-4$, AdamW as the optimizer, and Binary Cross Entropy as the loss function.

To train the model we solely relied on the training data provided by the organizers. For Subtask 2b, this dataset comprised 1200 instances, each consisting of a meme image paired with its corresponding text. The validation set contained 150 instances, while the development set included 300 instances. The final test sets encompassed 600 memes in English, 100 memes each in Bulgarian and North Macedonian, and 160 memes in Arabic. The distribution of labels across the training, validation, and development sets is presented in Table 2[7].

Label	Train	Val	Dev
propagandistic	800	100	200
non-propagandistic	400	50	100

Table 2: Label distribution on the training, validation, and development set for Subtask 2b.

All results were evaluated using the official classification measure adopted by the task organizers for Subtask 2b, i.e., Macro-F_1.

To select our best model, we performed a search over the best set of hyperparameters. Specifically, we varied the rank of the LoRA weight matrices (LoRA R), their alpha regularization factor (LoRA Alpha), the dropout rate of the LoRA weights (LoRA Dropout), and most importantly, we controlled the number of tokens by a Token Merging strategy (see Section 2.1).

4 Results

The task was organized into two main evaluation phases: a development phase, during which only training and unlabeled development data were accessible, and a test phase, wherein the gold labels for the development set were disclosed alongside the unlabeled test sets in four languages: English, Arabic, Bulgarian, and North Macedonian. In this section, we outline the results obtained by our experiments in both phases.

Development phase. In this phase, we conducted tests on the unlabeled development set, employing various combinations of image encoders and LLMs. For each combination, we set LoRA R and LoRA

[2] https://huggingface.co/openai/clip-vit-base-patch32

[3] https://huggingface.co/openai/clip-vit-large-patch14

[4] In particular `TinyLlama-1.1B-Chat-v1.0`: https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0

[5] https://huggingface.co/microsoft/phi-1_5

[6] https://huggingface.co/microsoft/phi-2

[7] For a comprehensive understanding of the dataset development and composition for each subtask, readers are encouraged to refer to the primary report of the task (Dimitrov et al., 2024).

Alpha to 8 and LoRA Dropout to 0.2. Notably, we omitted token merging during this phase based on preliminary results from the validation set, which indicated no significant performance enhancements with this setting. The results presented in Table 3 demonstrate that using larger Image Encoders, such as CLIP Large, yields an average increase of 0.7% in terms of Macro-F_1 performance. This enhancement may be attributed to higher-dimensional embeddings compared to their Base counterparts, even though it also produces a larger number of tokens due to a smaller patch size.

Image Encoder	LLM	Macro-F_1
CLIP Base	Phi-1.5	80.6
	Phi-2	80.6
	TinyLlama	80.8
CLIP Large	Phi-1.5	80.9
	Phi-2	81.6
	TinyLlama	81.6

Table 3: Macro-F_1 results on the development set of Subtask 2b across selected Image encoders and Large Language Models (for greater readability, F_1 scores are reported in percentage in all tables).

Among the various LLMs, both TinyLlama and Phi-2 exhibited identical performance. Consequently, we opted for TinyLlama and CLIP Large as the preferred models for further examination of model performance, using slightly adjusted hyperparameter settings. Specifically, we explored different numbers of tokens, LoRA ranks of 8, 16, and 32, and LoRA dropout values of 0.2, 0.3, and 0.4.

The results show that strong token merging strategies improve the model stability but limit its performance. The increase of the LoRA R greatly increases model instability due to the improved overfitting risk, while the increase in LoRA dropout greatly improves model performance, reaching the best Macro-F_1 result of 0.83487.[8] Our ablation study is depicted in Table 4.

Test phase. During the final evaluation phase of the campaign, we thus applied the best-performing setting described above on the test sets released by the task organizers. The results obtained are shown in Table 5. Overall, our team achieved reasonably good performance across both English and, albeit with a predictable decrease, in the zero-shot setting, where notable differences are observed. Upon comparing our performance with each top-ranked system in this subtask, we observe that the absolute difference between our system and the best-performing system in English (i.e., LMEME, which also ranks as the top system for Bulgarian) is 0.012, indicating that Themis achieved results very close to the top performer. For Bulgarian, the absolute difference is even smaller, at 0.003, suggesting that both systems exhibit very similar performance in this language. Conversely, for Arabic and North Macedonian, the difference is more pronounced, at 0.059 and 0.207, respectively, underscoring the limitations of our system in these languages.

[8] See the task leaderboard at https://propaganda.math.unipd.it/semeval2024task4/SemEval2024task4_dev.html

# tokens	LoRA R	LoRA Dropout	Macro-F_1
-	8	0.2	81.6
-	16	0.2	81.7
-	32	0.2	79.1
64	8	0.2	81.7
96	8	0.2	80.0
128	8	0.2	78.8
192	8	0.2	77.1
-	8	0.3	82.8
-	8	0.4	**83.4**

Table 4: Ablation study. In-depth results over the development set using CLIP Large and TinyLlama and different combinations of LoRA ranks (LoRA R column) and dropouts (LoRA Dropout column).

Language	Rank	Macro-F_1	Micro-F_1
English	5	79.8	82.6
Bulgarian	2	66.7	84.0
North Macedonian	8	47.9	72.0
Arabic	7	55.5	55.6

Table 5: Official results obtained across different languages on the test set.

5 Discussion and Error Analysis

Despite achieving promising results, in terms of Macro-F_1 scores, our model still occasionally misclassifies instances, particularly in cases involving the propagandistic nature of memes. Although the labeled test set was not made available by the organizers, we were still able to inspect more in detail the results obtained on the development set. Figures 3 and 4 illustrate examples of false positive and false negative predictions, respectively. In both

instances, we can formulate hypotheses regarding why and when our model generates errors. The false positive example may be misconstrued as employing a "Slogan" based persuasion technique, possibly due to text present in the sign. The false negative could stem from the model's inability to recognize inherent sarcasm due to limited sarcastic examples, further train with a larger dataset could mitigate this issue. The confusion matrix depicted in Figure 2 reveals an uneven distribution of errors across both classes, indicating a bias towards the propagandistic class. This bias could be attributed to the imbalance in the training set.

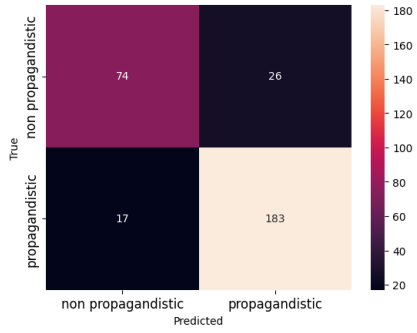

Figure 2: Confusion matrix of Themis predictions on the development set.

Figure 3: Example of false positive.

6 Conclusions

In this study, we introduced Themis, a novel model for analyzing multimodal memes by integrating LLMs and contrastive learning image encoders. Through comprehensive experiments, Themis demonstrated remarkable efficacy in detecting persuasion techniques within memes, achieving a notable F1 score of up to 83.4%. Our findings

Figure 4: Example of false negative.

underscore the critical role of meticulous model architecture design and hyperparameter optimization in meme analysis tasks. Notably, Themis presents a robust solution to combat societal challenges posed by biased content online, offering a promising avenue for mitigating the spread of misinformation and promoting digital discourse integrity.

Code availability

The code for our Themis model and train strategy is available on GitHub at: https://github.com/demonprin/Themis-SEMEVAL-public

Acknowledgements

The work has been partially supported by the project DEMON "Detect and Evaluate Manipulation of ONline information" funded by MIUR under the PRIN 2022 grant 2022BAXSPY (CUP F53D23004270006, NextGenerationEU), by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU (NextGenerationEU), by project eINS Ecosystem of Innovation for Next Generation Sardinia (CUP F53C22000430001) under the NRRP MUR program funded by the EU - NGEU (NextGenerationEU) and project NEST "Network 4 Energy Sustainable Transition–NEST" (CUP F53C22000770007) under the NRRP MUR program funded by the EU - NGEU (NextGenerationEU).

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. 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. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

Cédric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. 2022. Learning to merge tokens in vision transformers. *ArXiv*, abs/2202.12015.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model.