nowhash at SemEval-2024 Task 4: Exploiting Fusion of Transformers for Detecting Persuasion Techniques in Multilingual Memes

Abu Nowhash Chowdhury[1] **and Michal Ptaszynski**[2]
[1]Asian University for Women, Chattogram 4000, Bangladesh
[2]Kitami Institute of Technology, Kitami 090-8507, Japan
nowhash.chowdhury@auw.edu.bd and michal@mail.kitami-it.ac.jp

Abstract

Nowadays, memes are considered one of the most prominent forms of medium to disseminate information on social media. Memes are typically constructed in multilingual settings using visuals with texts. Sometimes people use memes to influence mass audiences through rhetorical and psychological techniques, such as causal oversimplification, name-calling, and smear. It is a challenging task to identify those techniques considering memes' multimodal characteristics. To address these challenges, SemEval-2024 Task 4 introduced a shared task focusing on detecting persuasion techniques in multilingual memes. This paper presents our participation in subtasks 1 and 2(b). We use a finetuned language-agnostic BERT sentence embedding (LaBSE) model to extract effective contextual features from meme text to address the challenge of identifying persuasion techniques in subtask 1. For subtask 2(b), We finetune the vision transformer and XLM-RoBERTa to extract effective contextual information from meme image and text data. Finally, we unify those features and employ a single feed-forward linear layer on top to obtain the prediction label. Experimental results on the SemEval 2024 Task 4 benchmark dataset manifested the potency of our proposed methods for subtasks 1 and 2(b).

1 Introduction

Modern social media represents a prominent environment to disseminate information to a vast community in real time. Hence, persuasion techniques are often embedded in social media content to subliminally influence people and their unconscious opinions. Such techniques are now incorporated in memes due to the increasing popularity among social media users. The visual aspect of memes adds to the effectiveness of grabbing people's attention than purely word-based messages. Manipulators and propagandists now treat it as an effective tool to promote and achieve their nefarious agendas. Sometimes different organizations use it to spread fake news or propaganda which causes social chaos and incitement of hate, which could result in harm or even human casualties. Hence, the detection of persuasion techniques embedded in memes appears as a formidable task to shield individuals from deceit. Moreover, detecting these techniques from memes is a challenging task since it requires a nuanced understanding of images, and texts, and a proper appreciation of the satirical characteristics of memes. To address these challenges, SemEval-2024 introduced a shared task focusing on detecting persuasion techniques from multilingual memes (Dimitrov et al., 2024). This task comprises three subtasks. Whereas the first task is based on identifying 20 persuasion techniques from meme texts. This is a hierarchical multilabel text classification task. Tasks 2(a) and 2(b) are based on multi-modal contents. Task 2(a) is a hierarchical multimodal multilabel classification task where the proposed system needs to identify 22 persuasion techniques from multimodal memes. Task 2(b) is a multimodal binary classification task where the participants need to apply the multimodal information expressed by memes to classify them into whether they contain a persuasion technique or not. A data sample of each task along with corresponding labels was articulated in Table 1.

However, some prior works have been done on identifying persuasion techniques from texts and visuals. SemEval 2023 shared task 3 introduced a subtask based on identifying persuasion techniques used in news articles (Piskorski et al., 2023). Most of the participants used different multilingual transformer models to tackle the challenge of this task. APatt (Purificato and Navigli, 2023) utilized an ensemble of different pre-trained transformer models e.g., XLNet, RoBERTa, BERT, ALBERT, and De-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 133–138
June 20-21, 2024 ©2024 Association for Computational Linguistics

Table 1: Sample Data of subtask 1, 2(a), and 2(b) of SemEval 2024 Task 4

Task No.	Sample Data	Label
Subtask 1	WHEN THE POWER OF LOVE IS GREATER THAN THE LOVE OF POWER, THE WORLD WILL KNOW PEACE	Loaded Language, Black-and-white Fallacy/Dictatorship, Slogans
Subtask 2(a)	Time To Straighten Out What Is Happening In Our Country! `prop_meme_4398.png`	Flag-waving,Glittering generalities (Virtue),Black-and-white Fallacy/Dictatorship
Subtask 2(b)	I MISSED THE SUPERBOWL-WHO WON?\\nEVERYONE WHO DIDN'T WATCH IT `prop_meme_4388.png`	non_propagandistic

BERTa incorporated by weighted average. Another team, KInITVeraAI (Hromadka et al., 2023) used fine-tuned XLM-RoBERTa-large model to address the multilingual characteristics of this task. They experimented with different prediction threshold values to find the optimal one.

To detect persuasion techniques in texts and images, SemEval 2021 Task 6 introduced three subtasks including multilabel text classification, span identification, and multi-modal multilabel classification task (Dimitrov et al., 2021). The top-performing team on the multilabel text classification task (Tian et al., 2021) leveraged five fine-tuned transformer models: BERT, RoBERTa, XLNet, DeBERTa, and ALBERT. They made use of external PTC corpus (Da San Martino et al., 2020) along with given training data to train these transformer models. Team NLPIITR (Gupta and Sharma, 2021) made use of a fine-tuned RoBERTa model to address the challenge of this task. Team Volta (Gupta et al., 2021) explored the potency of fine-tuned BERT and RoBERTa models for both multi-label text classification and span identification tasks and used RoBERTa Large for the final model. Their proposed architecture ranked top on span identification tasks. For the multi-label multimodal classification task, they tested the performance of the ensemble of multimodal transformers e.g., UNITER, VisualBERT, and LXMERT alongside unimodal transformers e.g., BERT, and RoBERTa. The winning team on subtask 3 (Feng et al., 2021) experimented with the ensemble of fine-tuned DeBERTa and ResNET, DeBERTa and BUTD, and ERNIE-ViL models to address the challenge of leveraging features from different data modalities.

In this paper, we demonstrate our proposed architecture to address the challenges of Subtask 1 (multi-label hierarchical text classification) and Subtask 2(b) (multi-modal binary classification) of SemEval 2024 Task 4. For subtask 1, we utilize a fine-tuned Language-agnostic BERT Sentence Embedding (LaBSE) model to extract effective contextual features of meme texts. Next, we utilize an ensemble of Vision Transformer and XLM-RoBERTa models to address the challenge of multilingual and multi-modal characteristics of subtask 2(b).

The remaining part of the manuscript is outlined as follows: The pictorial description of our proposed methods for both tasks is articulated in Section 2. Section 3 presents the experimental setup, result, and evaluation. We conclude this manuscript with some future research directions in Section 4.

2 Proposed Architecture

2.1 Subtask 1: Hierarchical Multi-label Persuasion Techniques Classification from Meme Text

The main objective of our system is to detect available persuasion techniques in meme text from 20 pre-defined persuasion technique categories. An overview of our proposed persuasion technique detection framework is shown in Figure 1.

Upon obtaining the meme texts, we employed Language-agnostic BERT sentence embedding (LaBSE) on top of Flair's Transformer Document Embeddings to generate effective document embedding vectors. Further, those document vectors are then fed to a single-layer feed-forward linear classifier to obtain the prediction label.

2.1.1 Language-agnostic BERT Sentence Embedding (LaBSE)

LaBSE is a multilingual transformer-based Language-agnostic BERT Sentence Embedding model developed by (Feng et al., 2020). It was trained on 6 Billion translation pairs and can generate sentence-level shared embedding features for 109 languages. To obtain optimal representations of multilingual sentences, LaBSE integrates

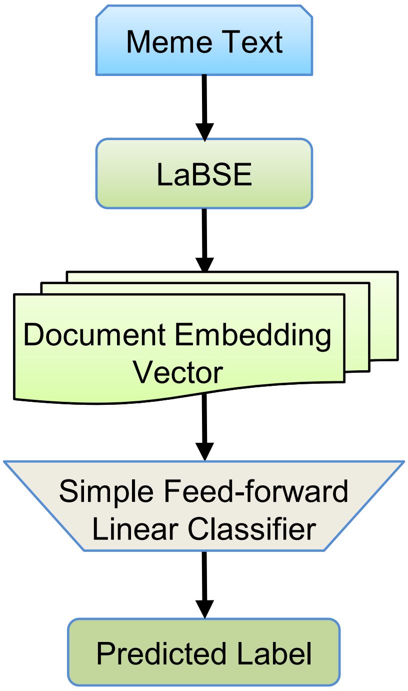
Meme Text
LaBSE
Document Embedding Vector
Simple Feed-forward Linear Classifier
Predicted Label

Figure 1: Proposed Framework of Subtask 1.

both monolingual and cross-lingual representations. It incorporates Multilingual BERT utilizing the masked language model and transformer language model with a translation ranking task alongside bidirectional dual encoders. We finetuned the LaBSE model on the benchmark dataset to capture the task-specific context effectively.

2.1.2 Transformer Document Embeddings

Document embedding represents embedding features of a full sentence rather than individual tokenized features. Flair's transformer document embeddings (Akbik et al., 2019) furnish an embedding for the entire text. We can extract embeddings directly from a pre-trained transformer model for a full sentence which enables us to capture the context of a sentence effectively. In our proposed architecture, we leverage the LaBSE model with transformer document embedding to obtain sentence-level embedding for a particular meme text.

2.2 Subtask 2: Multimodal Binary Classification Task

Figure 2 illustrates our proposed framework for subtask 2(b) where we tackled the challenges of multimodal meme classification.

Upon obtaining meme images and meme texts, we utilize a vision transformer and XLM-RoBERTa model to extract embedding features for both the meme images and meme texts. To tackle the multimodal characteristics of this task, we then integrated both visual and textual embedding features together and fed the integrated features to a single-layer feed-forward linear classifier to obtain the final prediction label.

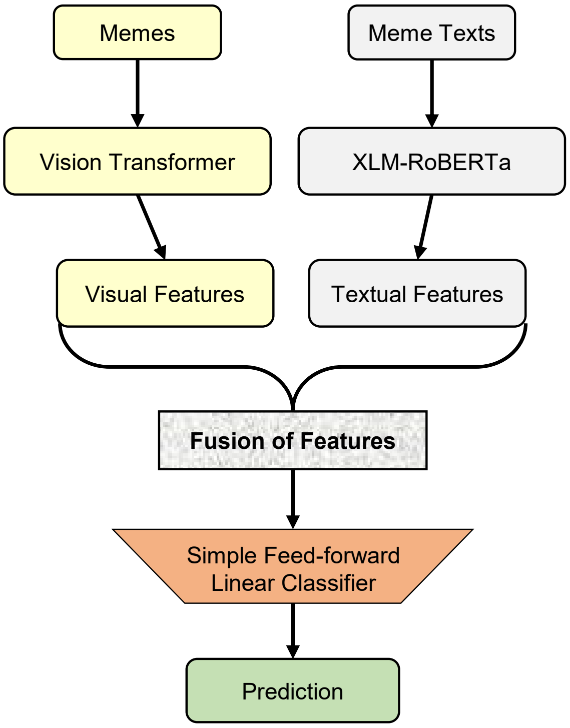

Figure 2: Proposed Framework of Subtask 2(b).

2.2.1 Vision Transformer

The vision transformer (ViT) is a self-supervised transformer encoder model pre-trained on a large image corpus (Dosovitskiy et al., 2020). ViT generates lower dimensional linear embedding by splitting the input image into fixed-size patches and flattening the patches. After adding positional embeddings, the flattened patches are fed into a standard transformer encoder as a sequence of tokens. The ViT encoder's internal architecture is similar to that of the original transformer. We utilize the finetuned ViT model facebook/dino-vitb16 checkpoint[1] (Caron et al., 2021) to extract effective visual information from memes.

2.2.2 XLM-RoBERTa

XLM-RoBERTa is a cross-lingual sentence encoder introduced by the Facebook AI group (Conneau et al., 2019). It was trained on a large 2.5 TB Common Crawl(CC) corpus containing over 100

[1] https://huggingface.co/facebook/dino-vitb16

languages. XLM-RoBERTa showed SOTA performance in various cross-lingual tasks (Eronen et al., 2022, 2023b,a). Both the base and large variants of XLM-RoBERTa contain 250M and 560M parameters, respectively with 250K vocabulary. In our proposed multimodal architecture, we utilized the finetuned XLM-RoBERTa large version to extract an effective representation of meme texts.

2.2.3 Fusion of Features

The fusion of high-level features from different data modalities in a neural architecture is conventional to tackle the challenge of representing multimodal features (Kumar and Nandakumar, 2022), (Pramanick et al., 2021), (Velioglu and Rose, 2020). In our proposed multimodal framework, we concatenate visual and textual features extracted from the finetuned ViT and the finetuned XLM-RoBERTa model for the effective representation of multimodal features.

2.3 Prediction Module

For both subtasks 1 and 2(b), We employed a single-layer feed-forward linear layer with SoftMax activation function to obtain the prediction, like in the equation 1 below.

$$q = Wp + b \tag{1}$$

Here, the input and output feature vectors are represented by p and q respectively. W is the weight matrix and b indicates the bias.

3 Experiments

3.1 Dataset Description

For subtasks 1 and 2(b), we utilized the dataset provided by the SemEval 2024 Task 4 organizers (Dimitrov et al., 2024) to train and finetune our proposed frameworks. Table 2 shows the detailed statistics of the dataset.

To evaluate the performance of our proposed frameworks, we utilized the hierarchical F1 score for subtask 1 and macro F1 score for subtask 2(b) as per the benchmark of SemEval 2024 Task 4 (Dimitrov et al., 2024).

3.2 Experimental Setup

We utilized the Google Colaboratory platform for system implementation, training, parameter tuning, and performance analysis. For subtask 1, we utilized the LaBSE model on Flair's NLP framework. The parameters used to train and finetune our model are illustrated in Table 3.

We made use of the vision transformer and XLM-RoBERTa model to tackle the challenge of subtask 2(b). The parameters used to train the vision transformer and XLM-RoBERTa are shown in Table 4 and Table 5, respectively.

Table 2: The statistics of the dataset.

Language	#Train	#Val	#Dev	#Test
Subtask 1:				
English	7000	500	1000	1500
Bulgarian	-	-	-	426
North Macedonian	-	-	-	259
Arabic	-	-	-	100
Subtask 2(b):				
English	1200	150	300	600
Bulgarian	-	-	-	100
North Macedonian	-	-	-	100
Arabic	-	-	-	160

Table 3: Optimal parameter settings for subtask 1.

Parameters List	Search Space	Value
Epochs	{4}	4
Batch size	{4}	4
Learning rate	{5e-5}	5e-5
Optimizer	{Adam, MADGRAD}	Adam
Multi-label Threshold	{0.1, 0.2, 0.30}	0.1

Table 4: Optimal parameter settings used in Vision Transformer.

Parameters List	Search Space	Value
Epochs	{4,6,8}	8
train_batch_size	{4,8,16}	8
eval_batch_size	{4,8,16}	8
Learning rate	{4e-5, 5e-5, 6e-5}	6e-5
Optimizer	{AdamW}	AdamW

Table 5: Optimal parameter settings used in XLM-RoBERTa.

Parameters List	Search Space	Value
Epochs	{4,6,8}	6
train_batch_size	{4,8,16}	4
eval_batch_size	{4,8,16}	4
Learning rate	{4e-5, 5e-5, 6e-5}	6e-5
Optimizer	{AdamW}	AdamW

Table 6: Synopsis of our proposed system performance in subtask 1.

Language	Hierarchical		
	F1 score	Precision	Recall
English	0.64096	0.61167	0.67320
Bulgarian	0.48627	0.46007	0.51563
North Macedonian	0.42558	0.41395	0.43788
Arabic	0.40370	0.35989	0.45965

Table 7: Synopsis of our proposed system performance in subtask 2(b)

Language	F1 Macro	F1 Micro
English	0.49845	0.51500
Bulgarian	0.43363	0.45000
North Macedonian	0.42857	0.52000
Arabic	0.49831	0.53125

3.3 Results and Analysis

In this section, we assess the performance of our submitted systems in SemEval 2024 Task 4 subtasks 1 and 2(b). The test dataset comprises four languages including English, Bulgarian, North Macedonian, and Arabic. Table 6 and Table 7 illustrate the performance of our model for subtasks 1 and 2(b), respectively.

For subtask 1, the experimental result shows that our proposed method achieved a good Hierarchical F1 score across the English, Bulgarian, North Macedonian, and Arabic datasets. We also report the hierarchical recall and hierarchical precision scores. This signifies the versatility of our approach across multiple languages. In subtask 2(b), there is still a significant performance gap between the top-performing systems and our system. One plausible reason might be the imbalanced fusion of visual and textual features.

4 Conclusion and Future Works

In this manuscript, we presented our proposed frameworks to address the challenge of SemEval 2024 Task 4 subtasks 1 and 2(b). We employed the LaBSE model to address the multilingual characteristics of subtask 1 whereas Vision Transformer and XLM-RoBERTa models were employed to address the multi-modal and multilingual characteristics of subtask 2(b). Both of our methods showed competitive performance over other participant's systems.

In the future, we aspire to explore the effectiveness of different multimodal transformer models' performance on this task. We also have a plan to exploit the external knowledge for a better understanding of memes for this task.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023a. Enhancing cross-lingual learning: Optimal transfer language selection with linguistic similarity. *Science Talks*, 6.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023b. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. 2022. Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management*, 59(4):102981.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at semeval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1075–1081, Online. Association for Computational Linguistics.

Vansh Gupta and Raksha Sharma. 2021. NLPIITR at SemEval-2021 task 6: RoBERTa model with data augmentation for persuasion techniques detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1061–1067, Online. Association for Computational Linguistics.

Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. KInITVeraAI at SemEval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 629–637, Toronto, Canada. Association for Computational Linguistics.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

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

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Antonio Purificato and Roberto Navigli. 2023. APatt at SemEval-2023 task 3: The sapienza NLP system for ensemble-based multilingual propaganda detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 382–388, Toronto, Canada. Association for Computational Linguistics.

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.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.