Puer at SemEval-2024 Task 4: Fine-tuning Pre-trained Language Models for Meme Persuasion Technique Detection

Jiaxu Dao, Zhuoying Li, Youbang Su, Wensheng Gong
School of Technology
Pu’er University
{daojiaxu, lizhuoying, suyoubang, gongwensheng}@peu.edu.cn

Abstract

The paper summarizes our research on multilingual detection of persuasion techniques in memes for the SemEval-2024 Task 4. Our work focused on English-Subtask 1, implemented based on a roberta-large pre-trained model provided by the transforms tool that was fine-tuned into a corpus of social media posts. Our method significantly outperforms the officially released baseline method, and ranked 7th in English-Subtask 1 for the test set. This paper also compares the performances of different deep learning model architectures, such as BERT, ALBERT, and XLM-RoBERTa, on multilingual detection of persuasion techniques in memes. The experimental source code covered in the paper will later be sourced from Github.

1 Introduction

Memes has been steadily increasing as human behavior as social media platforms have become more prevalent. This type of content is known for its rapid spread, achieved through the manipulation of audience psychology and the blurring of logical relationships.

Memes are generally made up of stacked images and text. The essence of its expression in order to generate an emotional effect is actually the skillful role of three persuasive strategies (Davison, 2012) in rhetorical portions:

1) Ethos: This involves the strategic employment of statements from individuals endowed with authority or credibility, thereby persuading the audience of the veracity of the content and augmenting its perceived legitimacy.

2) Pathos: By sharing personal anecdotes or experiences, memes forge a connection with the audience, evoking emotional resonance and deepening the affective engagement with the content.

3) Logos: The application of logical arguments and reasoning enhances the structural integrity and coherence of the message, fortifying its persuasiveness.

If we further split these three categories of persuasion strategies into twenty-two, scientists are able to obtain textual and visual features from memes for analysis. For instance, it is feasible to efficiently decrease or prevent the spread of hate speech, racial discrimination, and deceptive information by analysing memes, then simultaneously preserving the peace and stability of social media.

Memes can assist merchants in quickly capturing market trends, allowing them to carry out advertising and marketing operations more effectively and raise brand influence. Memes helps media workers in understanding the concerns of their audiences. Memes in politics have the potential to help voters demonstrate their policy views. The goal of the task is to classify corpora of text in memes and assign them to relevant persuasive strategies. Our work in SemEval-2024 Task 4 focuses on subtask 1,and this is a multi label classification task.

Our contrbutions can be highlighted as follows:

1) We explored new possibilities by screening models for news texts and multilingual corpus models. Fine-tuning using the social media posts corpus on the roberta-large model, and the experiment obtained hierarchical F1 of 0.647 on the English - Subtask 1 the dev set.

2) In SemEval-2024 Task 4, our model has an hierarchical F1 result of 0.66 in the English - Subtask 1 the test set, and our model ranks 7th on the leaderboard.

2 Related Work

Since the introduction of BERT in 2018(Devlin et al., 2018), its impact on the landscape of natural

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 64–69
June 20-21, 2024 ©2024 Association for Computational Linguistics

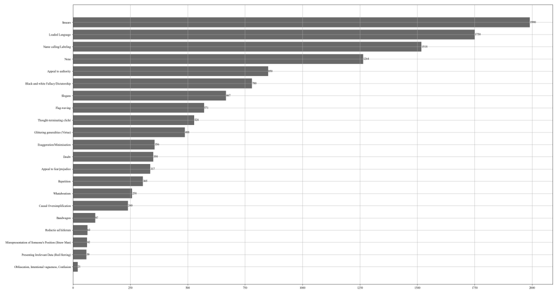


Figure 1: Number of Samples

ID	text	labels
67641	WHEN YOU'RE THE FBI, THEY LET YOU DO IT.	Thought-terminating cliché
66402	PUTIN'S SECRET CAMOUFLAGE ARMY	none
71251	Heaven has a Wall and strict immigration policies. Hell has open borders. President Donald J. Trump	Appeal to authority, Exaggeration/Minimisation
65282	ME VOTING ANTI-TRUMP IN 2016 ME VOTING ANTI-TRUMP IN 2020	Repetition

Table 1: Data Sample

language processing and multimodal analysis has been profound(Khare et al., 2021). BERT and its advanced derivatives, such as RoBERTa(Liu et al., 2019), XLM-RoBERTa(Xie et al., 2021), and ALBERT(Lan et al., 2019), have demonstrated their robust capabilities in a broad spectrum of applications, ranging from sentiment analysis to complex multimodal tasks that combine textual and visual data. Notably, RoBERTa has been recognized for its superior performance in accurately classifying sentiment(Liao et al., 2021), emotion(Kamath et al., 2022), and offensive content(Xu and Liu, 2023), highlighting the model's efficiency as a sophisticated text encoder.

The advent of these models has revolutionized the approach to analyzing diverse datasets and tasks, enabling nuanced understanding and processing of complex language patterns. This has been particularly evident in the domain of multimodal research, where BERT-based models have been instrumental in advancing the study of visual and textual data integration(Khan and Fu, 2021; He and Hu, 2021; Lee et al., 2021).

The success of these models in such a unique and culturally rich context exemplifies their broad applicability and the expanding frontiers of computational linguistics and content analysis. In conclusion, the inclusion of BERT and its variants in the analysis of persuasion techniques in memes marks a significant milestone in the field(Avvaru and Vobilisetty, 2020; Kougia and Pavlopoulos, 2021; Khedkar et al., 2022). It underscores the models' unparalleled flexibility and their emerging role in understanding the complexities of human communication in the digital age. As these models continue to evolve, their contribution to bridging the gap between textual and visual data analysis will undoubtedly pave the way for groundbreaking research and applications across diverse disciplines.

3 System Overview

3.1 Datasets

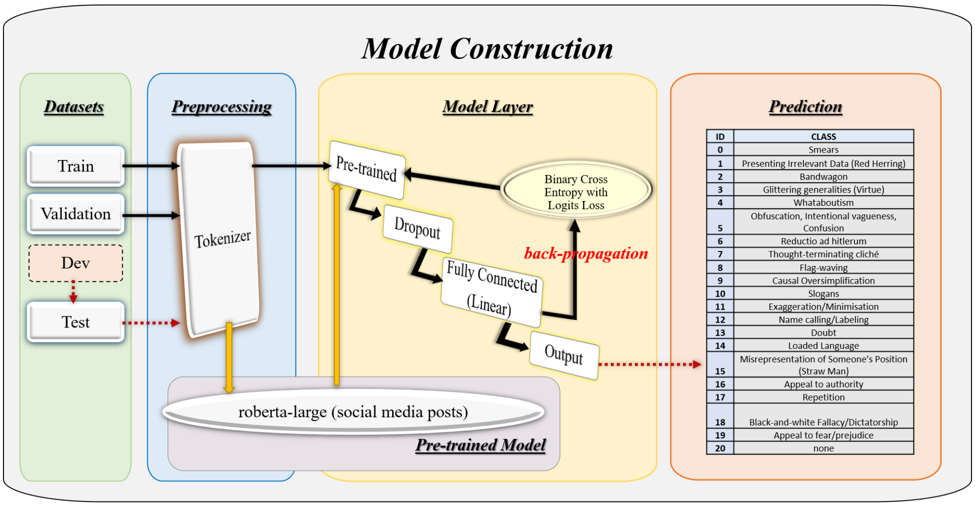
Our experiment employed four distinct datasets: the training set, validation set, development (dev) set, and test set, all formatted in JSON. The datasets feature a minimum sentence length of one. The training set comprises 7,000 entries, categorized into 20 distinct classes, showcasing an average sentence length of 19.94 and a maximum of 253. The validation set includes 500 entries, with an average sentence length of 18.85 and a maximum reaching 333. The development set, containing 1,000 samples, presents an average sentence length of 18.73 and a peak length of 145. Lastly, the test set encompasses 1,500 instances, with the sentences averaging 18 words in length.

Table 1 presents the sample dataset, illustrating the structured data used in our analysis.

Figure 1 sorts the distribution of categories in the training set in descending order of frequency, highlighting the frequency of each category. The term "None" denotes instances lacking specific classification. According to the depicted statistics, the category "Smear" constitutes the most significant portion of the dataset. In contrast, categories such as "Obfuscation", "Intentional vagueness" and "Confusion" represent the smallest proportions.

3.2 Pre-trained Model

The research team tends to choose from models related to news, tweets, and comments. The research team tested a number of models and deciding that Jochen Hartmann's sentiment-roberta-large-english-3-classes model (As shown in Table 2) while it received the best ratings. A comparison of outcomes from multiple models will be presented in the results section.



ID	CLASS
0	Smears
1	Presenting Irrelevant Data (Red Herring)
2	Bandwagon
3	Glittering generalities (Virtue)
4	Whataboutism
5	Obfuscation, Intentional vagueness, Confusion
6	Reductio ad hitlerum
7	Thought-terminating cliché
8	Flag-waving
9	Causal Oversimplification
10	Slogans
11	Exaggeration/Minimisation
12	Name calling/Labeling
13	Doubt
14	Loaded Language
15	Misrepresentation of Someone's Position (Straw Man)
16	Appeal to authority
17	Repetition
18	Black-and-white Fallacy/Dictatorship
19	Appeal to fear/prejudice
20	none

Figure 2: The architecture of model construction

ID	Model
1	bert-base-uncased
2	bert-base-multilingual-cased
3	albert-base
4	roberta-base
5	xlm-roberta-base
6	roberta-large
7	roberta-large(social media posts fine-tuned)

Table 2: Pre-trained Model

The sentiment-roberta-large-english-3-classes model (Hartmann et al., 2021) is trained based on tweets on social media platforms such as Twitter and Instagram, and includes text that is expected to include captions from the sender in the tweet image and comments from other observers. RoBERTa is used to construct the model. Achieving a hold out accuracy of 86.1 % , this model is used to evaluate user comments on posts and identify if the user is willing to buy a certain product. It demonstrates that the model has high robustness and a strong capacity to extract complicated text features.

3.3 Model Construction

In English-Subtask 1, to commence our experiment, we utilize the officially provided Train.json and Validation.json files as the training and validation datasets for supervised learning. Additionally, we assess subsequent results using the officially available dev dataset.

Secondly, we'll perform data preprocessing. The training and validation sets are fed into the Tokenizer, and the pre-trained model roberta-large(social media posts fine-tuned) is used for word segmentation and vectorization processing.

Following that, regarding model structure:

1) Input processing: Feed the pre-trained model with the processed token.

2) Dropout processing: Enter the dropout layer after model processing and set the inactivation probability to 0.1.

3) Linear fully connected layer: 1024 features are carried into the linear fully connected layer.

4) Loss function: For multi label classification jobs, Binary Cross Entropy With Logits Loss (Wang et al., 2022)serves as the loss function throughout the backpropagation gradient calculation procedure. BCEWithLogitsLoss comes with a sigmoid function that can convert predicted result values into probabilities, and can automatically handle numerical instability while preventing the sigmoid function from overflowing upwards or downwards (Yue et al., 2023).

Finally, the output layer is made up of 21 neurons, 20 of which are classified and one of which is none. The architecture of model construction is shown in Figure 2.

4 Experiment Setup

4.1 Evaluation Metrics

For English-Subtask 1, the participating systems are evaluated using standard evaluation metrics, including precision, recall, and hierarchical F1 scores.

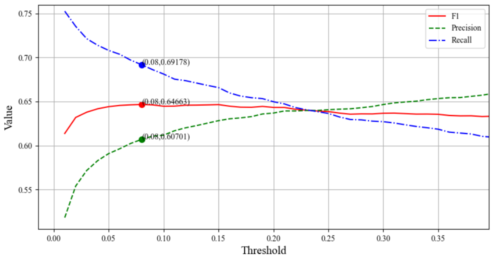


Figure 3: Impact of threshold on dev set

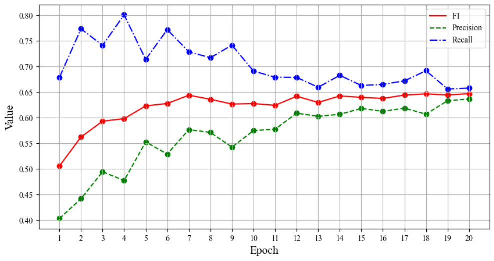


Figure 4: Impact of epoch on dev set

These metrics are calculated as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (1)$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

The organizers provided baseline models for each subtask. For English - Subtask 1, the Hierarchical F1 scores for the baseline model were 0.358 on the development set and 0.369 on the test set.

4.2 Threshold Selection

The experimental results in training tasks will depend on the threshold selection. We select the most optimal hierarchical F1 value for determining the threshold, assuming that recall and precision are of identical significance. With a 0.01 interval, the experiment increased the threshold from 0 to 1.

The red dots on the hierarchical F1 value curve in Figure 3 represent the experimental results, which show that the most suitable threshold value for hierarchical F1 value is approximately 0.08. In our threshold parameter experimentation, we attained a recall rate of 0.69 and a precision of 0.60. Owing to the threshold being established at 0.08, Figure 3 incorporates merely a fraction of the experimental data. The hierarchical F1 scores start to decline as the threshold surpasses 0.4.

4.3 Epoch Selection

The epoch was raised in the experiment from 1 to 20 at intervals of 1. The Figure 4 illustrates that the Precision is low and unstable and the Recall value is high but swings continuously when the epoch is under seven. As a result of the Precision and hierarchical F1 values' continued continuous increase, the experimental model's instability will grow. The Recall steadily stabilizes as the epoch gets closer to 20, while the hierarchical F1 value also tends to stabilize.

In addition to the above parameters, other training parameters are set in Table 3 below.

Params	Value
num_train_epochs	20
per_device_train_batch_size	4
per_device_eval_batch_size	8
warmup_steps	500
weight_decay	0.01
logging_steps	100
save_strategy	epoch
evaluation_strategy	epoch
learning_rate	$1.5e^{-5}$
threshold	0.08

Table 3: Training Arguments

5 Results

As Table 4 shown, the model's performance on the development set revealed an hierarchical F1 score of 0.64, a precision of 0.63, and a recall of 0.65. The results indicate that our model achieves better results than other models. The performance of English-Subtask 1 on the test set yielded an hierarchical F1 score of 0.66, a precision of 0.65, and a recall of 0.67, ultimately securing the 7th position in the ranking.

6 Conclusion

In our participation in SemEval-2024 Task 4, specifically English-Subtask 1, we focused on addressing the challenge of multi-label text classification. Our study investigated the impact of various pre-trained models on experimental outcomes and the influence of different hyperparameters on

Model	F1	Precision	Recall
bert-base-uncased	0.59335	0.60017	0.58668
bert-base-multilingual-cased	0.58840	0.58235	0.59459
albert-base	0.59484	0.58081	0.60957
roberta-base	0.62268	0.60781	0.63829
xlm-roberta-base	0.58612	0.57927	0.59313
roberta-large	0.63679	0.61831	0.65640
roberta-large (social media posts fine-tuned)	0.64708	0.63666	0.65786

Table 4: Dev Set Results

model performance. Ultimately, the adoption of the roberta-large model fine-tuned on social media posts led to outstanding performance, achieving a hierarchical F1 score of 0.66 on the test set and securing a commendable 7th position among English-Subtask 1 participants.

In our experimentation, we did not pursue a finer-grained classification within the multi-label task. Moving forward, our future research direction will pivot towards fine-grained multi-label classification. This would entail optimizing the loss function or implementing multi-level classification techniques to enhance the model's generalization capabilities.

Acknowledgements

This work was supported by the 2024 Science and Technology special project of Pu 'er University(PYKJZX202401 Research on medical relationship extraction task based on pre-trained language model). The authors would like to thank the anonymous reviewers for their constructive comments.

References

Adithya Avvaru and Sanath Vobilisetty. 2020. Bert at semeval-2020 task 8: Using bert to analyse meme emotions. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1094–1099.

Patrick Davison. 2012. The language of internet memes. *The social media reader*, pages 120–134.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. 2021. The power of brand selfies. *Journal of Marketing Research*, 58(6):1159–1177.

Jiaxuan He and Haifeng Hu. 2021. Mf-bert: Multimodal fusion in pre-trained bert for sentiment analysis. *IEEE Signal Processing Letters*, 29:454–458.

Rohan Kamath, Arpan Ghoshal, Sivaraman Eswaran, and Prasad Honnavalli. 2022. An enhanced context-based emotion detection model using roberta. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE.

Sujata Khedkar, Priya Karsi, Devansh Ahuja, and Anshul Bahrani. 2022. Hateful memes, offensive or non-offensive! In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2*, pages 609–621. Springer.

Vasiliki Kougia and John Pavlopoulos. 2021. Multimodal or text? retrieval or bert? benchmarking classifiers for the shared task on hateful memes. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 220–225.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Sanghyun Lee, David K Han, and Hanseok Ko. 2021. Multimodal emotion recognition fusion analysis adapting bert with heterogeneous feature unification. *IEEE Access*, 9:94557–94572.

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51:3522–3533.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xing Wang, Wenxian Yang, Bo Qin, Kexiang Wei, Yunyu Ma, and Daibing Zhang. 2022. Intelligent monitoring of photovoltaic panels based on infrared detection. *Energy Reports*, 8:5005–5015.

Shuyi Xie, Jian Ma, Haiqin Yang, Lianxin Jiang, Yang Mo, and Jianping Shen. 2021. Pali at semeval-2021 task 2: fine-tune xlm-roberta for word in context disambiguation. *arXiv preprint arXiv:2104.10375*.

Meijia Xu and Shuxian Liu. 2023. Rb_bg_mha: A roberta-based model with bi-gru and multi-head attention for chinese offensive language detection in social media. *Applied Sciences*, 13(19):11000.

Xiaohan Yue, Danfeng Liu, Liguo Wang, Jón Atli Benediktsson, Linghong Meng, and Lei Deng. 2023. Iesrgan: Enhanced u-net structured generative adversarial network for remote sensing image super-resolution reconstruction. *Remote Sensing*, 15(14):3490.