IUSTNLPLAB at SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes

Mohammad Osoolian
Iran University of Science and Technology
dsoolian@gmail.com

Erfan Moosavi Monazzah
Iran University of Science and Technology
moosavi_m@comp.iust.ac.ir

Sauleh Eetemadi
Iran University of Science and Technology
sauleh@iust.ac.ir

Abstract

This paper outlines our approach to SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes, specifically addressing subtask 1 in English language. The study focuses on model fine-tuning using language models, including BERT, GPT-2, and RoBERTa, with the experiment results demonstrating optimal performance with GPT-2. Our system submission achieved a competitive ranking of 17th out of 33 teams in subtask 1, showcasing the effectiveness of the employed methodology in the context of persuasive technique identification within meme texts.

1 Introduction

Propaganda is the term used when information is intentionally molded to promote a specific agenda. Memes typically involve combining an image with text. In deceptive memes, the image serves to either enhance or complement a technique employed in the text, or it independently conveys one or more persuasive techniques. In subtask 1 of SemEval-2024 Task 4 (Dimitrov et al., 2024), the challenge involves identifying which of the 20 persuasion techniques, organized hierarchically, are utilized, based on the textual content of a meme.

For this problem, GPT2 was chosen as the base model after experiments on GPT2, BERT and RoBERTa. After that, the model was fine-tuned on the given data set and after doing error analysis and comparing them with true labels, the threshold of sensitivity was changed manually to get best results. In addition, we tried to fine-tune model on SemEval-2023 Task 3 dataset which is similar to the given dataset for this task. However, the results didn't improve.

Regarding the noticeable change in scores just by changing the threshold of predicting labels, we realized the importance of error analysis and the easy tricks comes after actually understanding the behavior of model and it's problems.

We have made all the code necessary to replicate our results available in the paper's GitHub repository.[1]

2 Background

2.1 Dataset Description

The dataset consists of 7000 samples for training and 500 samples for validation. each sample contains three fields:

- id: A unique identifier assigned to each sample, facilitating the association with the corresponding meme image. It is noteworthy that, for the purposes of Subtask 1, the visual components of the memes, indicated by these IDs, are not considered in the training.
- text: this field is the textual content of the meme, as a single UTF-8 string. While the text is first extracted automatically from the meme, it has been post-processed to remove errors and formated in such a way that each sentence is on a single row and blocks of text in different areas of the image are separated by a blank row.
- label: it is a list of valid technique names used in the text. There are 22 techniques in this dataset which are leaf nodes of the hierarchy of persuasion techniques shown in Figure 1. However, only 20 of them are used for subtask 1.
- link: This field contains the social network link associated with the meme. It is imperative to acknowledge that certain samples may lack a corresponding link. In such cases, the term "null" is employed in lieu of a link.

You can see an example of training samples in Figure 2.

[1] https://github.com/mohammad-osoolian/SemEval-2024_task4

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1092–1096

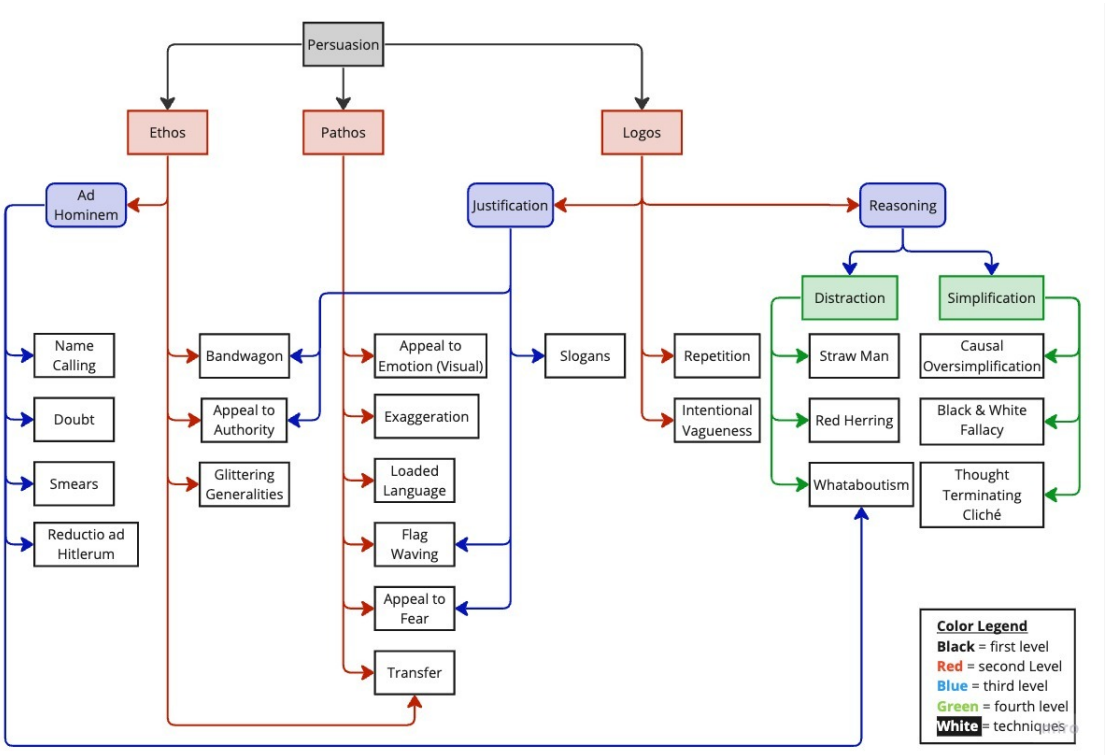

Figure 1: Hierarchy of persuasion techniques (Dimitrov et al., 2024)


```
{
"id": "66730",
"text": "WHEN THE POWER OF LOVE IS GREATER THAN THE
         LOVE OF POWER, THE WORLD WILL KNOW PEACE",
"labels": [
           "Loaded Language",
           "Black-and-white Fallacy/Dictatorship",
           "Slogans"
           ],
"link": "null"
},
```


Figure 2: An example in the training set

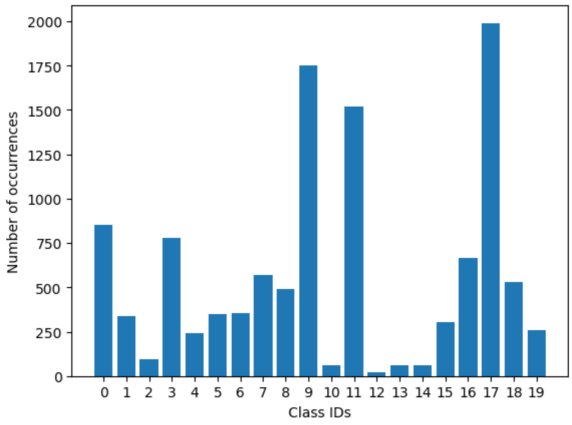

Figure 3: Number of Class occurrences in the labels

The distribution of classes in train data is shown in the Figure 3 and class names are shown in the Table 1.

2.2 Related Works

Prior to the SemEval 2024 event task, researchers have endeavored to address analogous challenges, contributing to the evolution of methodologies for detecting persuasive techniques in multimodal content.

The article "Detecting Propaganda Techniques in Memes" (Dimitrov et al., 2021) establishes a novel multi-label, multimodal task of automatically detecting propaganda techniques in memes. creating a dataset of 950 annotated memes covering 22 propaganda techniques, the authors provide a crucial resource for training and evaluating future detection models. In addition, by creating a dataset of 950 annotated memes covering 22 propaganda techniques, the authors provide a crucial resource for training and evaluating future detection models.

The article "SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup" (Piskorski et al., 2023) provides a publicly available dataset of annotated news articles, along with code and evaluation metrics. These resources serve as a valuable starting point for future research and development in multilingual news analysis tasks.

Number	Fallacy
0	Appeal to authority
1	Appeal to fear/prejudice
2	Bandwagon
3	Black-and-white Fallacy/Dictatorship
4	Causal Oversimplification
5	Doubt
6	Exaggeration/Minimisation
7	Flag-waving
8	Glittering generalities (Virtue)
9	Loaded Language
10	Misrepresentation of Someone's Position (Straw Man)
11	Name calling/Labeling
12	Obfuscation, Intentional vagueness, Confusion
13	Presenting Irrelevant Data (Red Herring)
14	Reductio ad hitlerum
15	Repetition
16	Slogans
17	Smears
18	Thought-terminating cliché
19	Whataboutism

Table 1: Class names and their IDs

2.3 Task evaluation and ranking

The hierarchical taxonomy of labels in this task necessitates a nuanced approach to evaluation. According to the task description, when predicting the ancestor node of a technique, only a partial reward is assigned, highlighting the hierarchical multilabel classification nature of the problem at hand.

To assess the performance of submissions, the chosen metric is the hierarchical F1 score (Kiritchenko et al., 2006). Hierarchical f1 score is a way of adapting the F1 score metric to be used for classification tasks with hierarchical structures. These structures involve classes having parent-child relationships, forming a kind of tree-like organization. It is crucial to note that the conventional F1 score is designed for flat classifications devoid of hierarchical relationships, making the hierarchical F1 score a pertinent choice for the evaluation of this task.

3 System overview

3.1 Model Architecture

Initially, our approach involved the utilization of three distinct models: BERT, RoBERTa, and GPT-2, all of which were subjected to fine-tuning on the training set. Subsequent evaluation based on the metrics outlined earlier revealed that the performance of the GPT-2 model surpassed that of its counterparts. (Table 2)

Given the superior performance observed with the GPT-2 model, we proceeded with this architecture for further refinement. The fine-tuning process ensued, culminating in the generation of our final results, which were subsequently submitted utilizing the GPT-2 model.

3.2 Fine tuning on extra dataset

Following the initial training on the provided dataset, our exploration extended to leveraging comparable datasets from previous studies and SemEval events. The SemEval-2023 Task 3 dataset, encompassing paragraphs extracted from diverse news articles and publications annotated with 19 distinct propaganda techniques, emerged as a pertinent source for augmenting our training data.

To ensure compatibility and coherence between the SemEval-2023 Task 3 dataset and our specific task dataset, a meticulous data cleaning process was undertaken. This involved the removal of uncommon tags, resulting in a curated dataset comprising 3,445 new samples. This augmented dataset was then incorporated into the fine-tuning phase of our model, aiming to enhance its adaptability and robustness across diverse text corpora.

3.3 Adjusting the prediction threshold

The model generates continuous probability values reflecting the likelihood of the presence of various persuasion techniques within the input text, rather than providing explicit binary predictions. A threshold is applied to discretize these probability values, where a value exceeding the threshold results in a prediction of 1, and otherwise, it is predicted as 0. The adjustment of this threshold played a pivotal role in refining the model's output, leading to noticeable improvements in performance.

In fine-tuning the threshold value, we tested a range of thresholds on the training set and assessed their performance using the F1-score. Based on the results depicted in Figure 4, we settled on a threshold value of 0.19.

Model	Accuracy	Precision Macro AVG	Recall Macro AVG	F1-Score Macro AVG
BERT	0.218	0.403	0.201	0.238
RoBERTa	0.232	0.344	0.179	0.2145
GPT2-medium	0.382	0.637	0.423	0.489

Table 2: Evaluation Metrics for GPT2, BERT and RoBERTa models on validation set

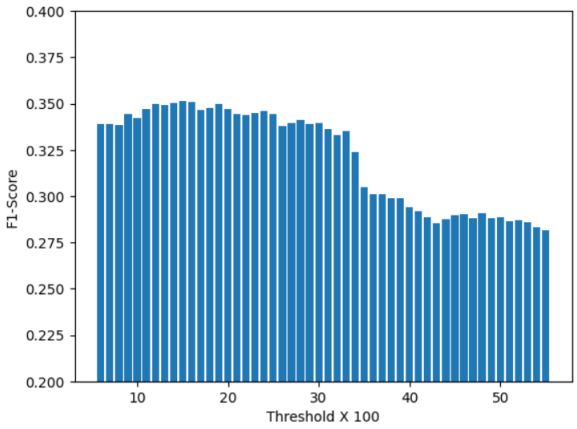

Figure 4: Adjusting the prediction threshold with F1-Score

4 Experimental setup

4.1 Dataset Split

The dataset is partitioned into distinct sets to facilitate comprehensive training, validation, and evaluation processes. The training set comprises 7,000 samples, while the validation set consists of 500 samples. The test set, used for submitting predictions, encompasses 1,500 samples.

Additionally, there exists a dev set that initially lacked labels and was subsequently annotated. This set was not incorporated into the model training process but only used to measure improvements in results and scores.

For the purpose of augmenting the training data, an extra dataset derived from SemEval-2023 Task 3 was considered. Following data cleaning, this supplementary dataset yielded 3,445 samples. However, despite this effort, training the model with the extra dataset did not yield discernible improvements. Consequently, the submitted model was not fine-tuned using this additional dataset.

4.2 Preprocessing Dataset

In the preprocessing of the main dataset, the initial step involved converting the data from JSON format to a tab-separated values (tsv) format. During this transformation, the "link" field in the samples was removed. The resulting dataset comprises columns for ID, Label, and Text.

As for the extra dataset, the samples were initially distributed across various files as paragraphs, with labels stored separately in different files. To align with the structure of the main dataset, each file was processed by splitting it into individual paragraphs. Subsequently, labels were gathered, and each paragraph was transformed into a unified sample with an assigned ID. To ensure compatibility, samples with labels not present in the main dataset were excluded, streamlining the integration of the extra dataset into the training process.

4.3 Evaluation Metrics

The system we have designed, only predicts the actual 20 classes which are the leaf nodes in the hierarchy of persuasion techniques. Therefore we have not used proposed hierarchical f1 score. The metrics we have used for our own evaluations are as follows:

- precision: Calculated individually for each class and expressed as total precision with macro averaging between classes. Precision serves to measure the accuracy of positive predictions.
- recall: Computed for each class and represented as total recall with macro averaging between classes. Recall measures the completeness of positive predictions.
- f1-score: Determined for each class and presented as the total F1 score with macro averaging. The F1-score serves as a comprehensive metric in classification tasks, considering both precision and recall.

5 Results

5.1 Overall Performance

Finally our model reached hierarchical f1-score of 0.624 and hierarchical precision of 0.631 and hierarchical recall of 0.617 in English language. In comparison, the baseline metrics for these categories were significantly lower at 0.368, 0.477, and 0.300, respectively. (Table 3)

Model	Hierarchical F1-score	Hierarchical Precision	Hierarchical Recall
First team	0.752	0.684	0.835
Our model (17th team)	0.624	0.631	0.617
Baseline	0.368	0.477	0.300

Table 3: Team ranking and model hierarchical scores

5.2 Analysis model predictions

By comparing the obtained results with the results of the first team, we see that the precision values are not much different, but the recall value for the first group is much higher than the recall of our model. This disparity indicates that our model may lack sensitivity and we can achieve better results by focusing on improving recall in the model.

6 Conclusion

In this paper, we examined different models and finally by choosing GPT2, we presented a model for the problem of identifying persuasion techniques in English memes. With the help of the presented model and adjusting the threshold for this model, we were able to reach a score of 0.624 for f1-score. Our work demonstrates the effect of choosing the appropriate model for training and the need to perform error analysis to improve the accuracy of the model.

References

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.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Advances in Artificial Intelligence*, pages 395–406, Berlin, Heidelberg. Springer Berlin Heidelberg.

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 multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.