Naming Persuasion Techniques in Political Memes using LLMs

Madeline Ou
mro2131

Isabel Snyder
irs2122

Abstract

The goal of our project is to evaluating the effectiveness of LLMs for detecting propaganda techniques in memes. This was a multilabel classification task that required models to predict the propaganda techniques present within the meme's textual data. Through fine-tuning and evaluating BERT-based models RoBERTa and DistilBERT, we found that both models performed significantly lower than expected, with F1 scores of roughly 0.4 to 0.5. This poor performance could be attributed to a variety of reasons from data quality to the way we implemented the models, so further work is necessary.

1 Introduction (15 pts)

For our project, we chose to work on the SemEval 2024 Task 4: "Multilingual Detection of Persuasion Techniques in Memes". We decided to limit our scope to just Subtask 1, which requires classifying the persuasion technique(s) used in a political meme using only the text appearing in the meme (the "textual context"). Our goal was to achieve an F1 score of 0.60 by fine-tuning a large language model (LLM) to solve the classification task.

Memes are an increasingly prevalent way of communicating, expressing beliefs, and connecting, especially among young people. The ease with which they are created and deciphered allows them to rapidly proliferate and spread. These properties have helped memes engender connection and shared culture, but also allow for large-scale disinformation and support small-scale social media echo chambers, whereby a group's beliefs are continuously reinforced by the media they consume and produce. Like other forms of media, they reflect the implicit and explicit biases of their authors. As memes continue to play an important role in developing modern culture and socializing our youth, work that contributes to better understanding their motives and effects is highly important.

SemEval Task 4 specifically focuses on memes created for propaganda, which the organizers define as "information ... purposefully shaped to foster a predetermined agenda" (Dimitrov et al., 2023). Memes can be an effective tool for such campaigns, quickly disseminating the views of the campaign's creators in a consumable format. Disinformation such as these memes is a likely factor in the US's increasingly polarized and polemic political sphere. Thus, identifying memes with a motive of propaganda can be incredibly important in recognizing how certain ideologies are spreading and containing them, if necessary.

By applying and fine-tuning LLMs to this task, we hope to explore the role of LLMs in society as harm-reducing agents while putting into practice some of the concepts we've learned in class, such as modern LLM architectures and attention. So, our primary research question is: To what extent are LLMs an effective tool for accurately detecting persuasion techniques in memes?

From a technical perspective, LLMs have been well-established as excellent text classifiers for sentiment or topic classification. However, this task is particularly difficult: memes often use very short amounts of text and rely on the visual impact of the image they display as well as the text (which subtask 1 ignores), persuasion techniques rely on subtlety, and any number of up to 20 persuasion techniques could apply, including none. As seen on the task website, the winning team for subtask 1 ended up with an F1 score of only 0.75247, the best of 40 teams (Dimitrov et al., 2023). Thus, we wanted to contribute to this problem with our own exploration into the efficacy of LLMs for solving this task.

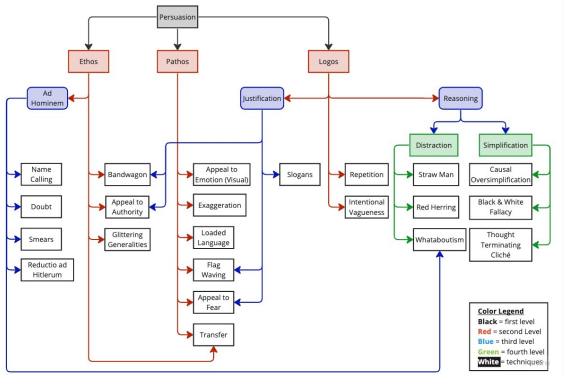

Figure 1: The persuasion technique hierarchy given in the SemEval Task 4. "Appeal to Emotion (Visual)" and "Transfer" are excluded for Subtask 1, since they rely on visual cues not present in the meme descriptions.

2 Methods (40 pts)

2.1 Datasets (10 pts)

To train and test our models, we used the SemEval Task 4 dataset provided by the task organizers for Subtask 1 (Dimitrov et al., 2023). All data collected for the purpose of this task was obtained by scraping public Facebook groups focusing on politics, vaccines, Covid-19, gender equality, and the Russo-Ukrainian war. It is a novel dataset that the organizers created. For Subtask 1, specifically, the data consists of descriptions of memes, rather than memes themselves. Each example is in the same form as this example:

```
{
    "id": "125",
    "text": "I HATE TRUMP\n\nMOST TER-
            RORIST DO",
    "labels": [
        "Loaded Language",
        "Name calling/Labeling"
        ],
    "link": "https://..."
}
```

The organizers provide 7,000 examples for the training set, 500 examples for the validation set, and 1,500 examples for the test set, each annotated with the applicable labeling of the given persuasion technique(s). The persuasion techniques examined in this task are a set of 20 labels organized in a hierarchical format, as can be seen above in Figure 1.

Descriptions and examples for each of the persuasion techniques can be found on this webpage, also linked on the homepage for SemEval Task 4 (Dimitrov et al., 2023). Among the given datasets for Subtask 1, the breakdown of examples by label can be seen in Table 1, in the second column of this page.

The text was already cleaned for us, so data pre-processing was relatively simple. However, there were many examples to which no labels applied. In the execution of our code, these examples were causing some errors, so we decided to apply a label of "None" to all examples without any labels in the pre-processing phase.

The data does include input texts of varied lengths, the longest including 2,334 characters and the minimum having no text at all. The mean input length is 117 characters. Similarly, the number of labels is varied, with the maximum having 8. The mean number of labels is 1.8.

2.2 Models/Approach (10 pts)

To evaluate the performance of our SemEval task, we selected two BERT-based models: RoBERTa and DistilBERT.

We decided to specifically focus on the BERT-based models because of their transformer-based architecture and self-attention mechanism. Both characteristics make the models highly effective and accurate at text classification tasks, as they are capable of handling more complex dependencies and relationships between words.

Additionally, because BERT-based models treat each label as a separate binary classification task, during training, the model learns to predict the presence or absence of each label independently. This structure marks BERT-based models extremely effective at multi-label classification, which was the specifics of our task.

2.2.1 RoBERTa: Robustly optimized BERT pretraining approach

RoBERTa is an optimized variant of the original BERT model, focusing on improving original BERT pretraining procedure by using more data and removing Next Sentence Prediction (NSP) objective (Liu et al., 2019). RoBERTa's size in parameters is 125 million, an incredibly large amount of learnable individual weights that increases training time but also increases the model's capacity for learning more complex relationships. Comparatively, BERT-base has roughly 110 million parameters. Additionally, RoBERTa is trained on a much

Label	Train	Validation	Test	Total
Appeal to authority	850	63	136	1049
Appeal to fear/prejudice	337	27	66	430
Bandwagon	97	7	16	120
Black-and-white Fallacy/ Dictatorship	780	53	98	931
Causal Oversimplification	240	21	53	314
Doubt	350	24	45	419
Exaggeration/ Minimisation	356	27	62	445
Flag-waving	571	42	89	702
Glittering generalities (Virtue)	488	36	71	595
Loaded Language	1750	135	303	2188
Misrepresentation of Someone's Position (Straw Man)	62	4	10	76
Name calling/ Labeling	1518	116	262	1896
Obfuscation, Intentional vagueness, Confusion	21	2	8	31
Presenting Irrelevant Data (Red Herring)	59	4	10	73
Reductio ad hitlerum	63	4	11	78
Repetition	305	23	46	374
Slogans	667	50	111	828
Smears	1990	142	282	2414
Thought-terminating cliché	528	38	78	644
Whataboutism	258	21	52	331
None	1264	88	156	1508

Table 1: Number of examples of each persuasion technique in each subset of the data.

larger corpus and with dynamic masking. As a result, RoBERTa has been shown to outperform BERT on several NLP tasks, including classification – making it a strong choice for our multilabel classification task. The specific checkpoint we used was roberta-base.

2.2.2 DistilBERT: Distilled version of BERT

DistilBERT is a smaller, faster, and lighter variant of original BERT model (Sanh et al., 2019) It's based on the concept of knowledge distillation, which involves training a smaller student model to approximate the performance of a larger teacher model. DistilBERT retains about 97% of BERT's performance while being 60% smaller and 60% faster. Thus, DistilBERT's size in parameters is roughly 66 million, and while the transformer architecture is the same, it was fewer layers – 6 layers instead of 12 layers. The specific checkpoint we used was distilbert-base-uncased.

Ultimately, this decision to go with BERT-based models was also based on the publicly available submissions for the task, with most relying on one or multiple versions of BERT.

2.3 Experiments (20 pts)

The task at hand is a multilabel classification problem, where each input meme text needs to be classified into one or more of the 20 propaganda techniques. The goal of the task was to have the model predict the correct set of fallacies for each meme text. Notably, due to constraints, we decided to predict only leaf nodes, evaluating the models based on pure correctness rather than awarding partial credit.

The data used were the training, validation, and test datasets provided by the task organizers. All datasets were a collection of meme text data, with only the training and validation datasets having the corresponding labels of various propaganda techniques given. The test dataset's meme text was unlabeled, used only during the final evaluation to provide performance metrics in comparing the different model configurations: RoBERTa and DistilBERT.

In our approach, we first preprocessed all the data. Using the model's respective tokenizers from the transformers library, we loaded and tokenized the meme text data. Then, 20 labels for the different propaganda techniques were encoded using the MultiLabelBinarizer from scikit-learn. This method transformed the list of techniques for each

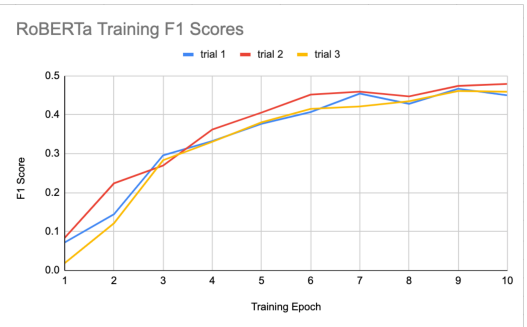

Figure 2: RoBERTa F1 scores on the validation data over ten epochs of training

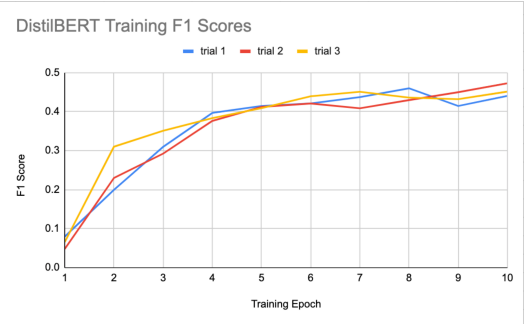

Figure 3: DistilBERT F1 scores on the validation data over ten epochs of training

text into a binary vector – where each entry corresponded to propoganda technique, and each value indicated whether it was present or not.

We used the pretrained RoBERTa and DistilBERT models with a sequence classification head, as the model was already fine-tuned for classification tasks. Our next job was specifically to fine-tune them to multilabel classification tasks. Fine-tuning happened over a process of 10 epochs. And during this process, loss was calculated with the BCEWithLogitsLoss for the loss function, as the model would output a probability for each label independently – suitable for multilabel tasks. Additionally, we used the AdamW optimizer, with a learning rate of 5e-5 and epsilon value of 1e-8, updating the models' weights during training based on loss gradient. Notably, all hyperparameters used were the default ones, with the exception of a batch size of 32. This was because training and fine-tuning a model as large as RoBERTa proved to be too intensive, as our GPU kept running out of memory.

During the training process, the models were evaluated on the validation set after each epoch using F1 score as the primary evaluation metric. Specifically, the F1 micro score was used. This metric is particularly useful for multilabel classification tasks where labels are not mutually exclusive because it calculates a single score by averaging per-label F1 scores. After training, the models were tested on the test dataset to assess its generalization ability, with F1 score being used again for evaluation.

For each model, this process was repeated 3 times, allowing us to better compare the results across both configurations.

3 Results/Analyses (15 pts)

During training, both models started with extremely low F1 scores, less than 0.1. Over the course of training, they first improved pretty quickly but started to level out around epoch 7 for RoBERTa and around epoch 5 for DistilBERT, seeing only marginal improvement after those points. These effects can be seen above in Figures 1 and 2 for RoBERTa and DistilBERT, respectively, which consist of plots of the validation F1 score over ten epochs of training. Ultimately, both ended at a point almost midway between 0.4 and 0.5 over the epochs of training.

During the final evaluation of the test data, the models performed pretty similarly, for which you can see the full results in Table 2. On average, RoBERTa performed a bit better on the test data: the mean F1 score is 0.4681 for RoBERTa and 0.4588 for DistilBERT.

model	**trial 1**	**trial 2**	**trial 3**	**mean**
RoBERTa	0.4832	0.4810	0.4410	**0.4681**
DistilBERT	0.4667	0.4590	0.4507	**0.4588**

Table 2: F1 score on the test dataset for each model across each trial.

These results were not what we had hoped to achieve, falling significantly below the 0.60 F1 score we were aiming for. There are a couple reasons this is likely the case: we created a "None" label for samples in the data with no assigned label, the data itself has imperfections, and limitations of time and Colab use inhibited our ability to tune the hyperparameters. Here, we examine each one of these factors.

First, creaing a new label for data with no label likely disadvantaged the models because predictions where all logits fell below our positive pre-

diction cutoff of 0.5 then failed to correctly assign *any* of the labels. However, in the original dataset, not assigning a label at all is also a correct answer in many cases. Therefore, our models likely would have had somewhat better performance if we had figured out a way to support model predictions of no label instead of creating a new label for memes that fell in that bucket.

Second, the data has imperfections and quirks that increase the difficulty of the task. As stated in Section 2.1 above, the longest meme description had over 2,000 characters, significantly above our truncation limit of 512 for the tokenizer. Similarly, at least one meme had no text description at all! Both of these present challenges in creating a good classifier and likely compound the issue mentioned above, though they only apply to a minority of samples in the training data.

Third, we were relying on Google Colab to quickly train and evaluate our models. However, Colab often disconnects from the runtime and ultimately has GPU usage limits for free accounts like ours. These problems were relatively minor, but solving them required time, our more limited resource. We planned on fine-tuning the hyperparameters themselves for the model, but didn't end up having the time to run multiple trials with differing values. The exception to this is `batch_size`, which we had initially set to 64 but changed to 32 for the training and validation Dataloaders. We ran out of available memory in our allocation of Colab when using the higher `batch_size` value, so we adjusted it downwards to fix the issue. As we will discuss below, there is more potential here that we would have liked to explore.

4 Related Work (15 pts)

Recent research on detecting persuasion techniques in text has gained considerable attention, especially in the context of analyzing discourse in online spaces, social media, and political content. For example, previous SemEval 2017 Task 11: "Fine-Grained Propoganda in News Articles" (Da San-Martino et al., 2019) is the first work to study propaganda at the fine-grain level: rather than labeling texts as propaganda or not, specific spans within the texts are analyzed and labeled with a corresponding specific propaganda technique out of 18 total techniques, such as "loaded language" or "appeal to authority." This work was foundational in building the tools for detecting various type of propaganda techniques, and it highlighted the challenge of recognizing subtle persuasive cues in langauge. The emphasis on this granular analysis functions as the foundational ground for our task to detect different persuasion techniques in memes.

Building upon this work, the SemEval 2023 Task 3: "Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multilingual Set up" aimed detect propaganda techniques in new articles covering 9 languages (Piskorski et al., 2023). This task significantly expanded the task of propaganda detection to a more global context, adding additional layers of complexity and challenges as cross-linguistic variability and cultural differences were introduces to the equation of persuasive communication. While our project does not tackle multilingual data, this research underscores the importance of adaptability and nuanced understanding, which is relevant for detecting diverse meme content.

Additionally, memes are inherently multimodal, combining images with text to create persuasive messages. Research into multimodal systems has grown in recent years, as traditional text-based analysis methods fail to account for the visual components that often play a significant role in persuasion. Because this content can be potentially used for harm, several studies have explored the use of prompting strategies for things such as hateful meme classification, aiming for detection from both text and visual modalities (Prakash et al., 2023). These studies utilize novel multimodal prompt-based model design to learn topics from both text and visual modalities by leveraging the language modeling capabilities of large language models, effectively extracting and clustering topics in memes to consider their semantic relations between text and visual (Prakash et al., 2023). Although this multimodal approach closely parallels the methodology of our project, where we fine-tune models like RoBERTa and DistilBERT, it differs by focusing exclusively on the text modality of memes rather than integrating visual analysis.

5 Conclusion (10 pts)

The primary goal of this project was to explore the effectiveness of large language models, specifically RoBERTa and DistilBERT, in detecting persuasion techniques in memes as part of the SemEval 2024 Task 4 Subtask 1. The project aimed to answer the following research question: To what extend

are LLMs effective tools for accurately detecting persuasion techniques in memes?

As we found, the single instances of RoBERTa and DistilBERT that we fine-tuned to our task were relatively ineffective, missing our goal of an F1 score of 0.60 on the test dataset. We did perform better than the task baseline value of 0.3687, but poorly on the whole. Our results would indicate that LLMs may not be effective at solving this task, but our project had serious limitations in its implementation, as discussed in Section 3. RoBERTa and DistilBERT are very similar models, each implementing the base BERT architecture. Whereas RoBERTa is a more robustly-trained version of BERT that takes advantage of hyperparameter optimization to create a better model, DistilBERT reduces the number of parameters in BERT while still maintaining its effectiveness. Our results seem to show that the small differences between the number of parameters and robustness of RoBERTa and DistilBERT have similarly small effects when fine-tuned for our dataset. To greater explore our research question, further research should examine a wider array of LLMs.

(However, the published team results on the task webpage shows that LLMs do face significant challenges in this domain, at least using this dataset: the winning team achieved the highest F1 score of all, at only 0.75 (Dimitrov et al., 2023).)

Throughout the project, we discovered many limitations. At the top level, the task is just extremely difficult for everyone – when we took a look at the data and tried annotating it ourselves, we found that it was tricky to correctly identify which exactly which persuasion techniques were present. It often feels like a judgment call instead of a clear-cut answer. Thus, the quality and consistency of the annotated data is itself a concern. In addition, the dataset source was solely comprised of Facebook memes, limiting its generalization as social media platforms have varied user bases, cultures, and meme formats that might not have been present. Additionally, while the task of identifying persuasive techniques in memes is cool, we found it hard to image it having any significant real-world impact or utility beyond specific academic or social research contexts.

Because it was the first time both of us were working with LLMs in such a capacity, we found it tricky to do even simply things like making sure the way we were preprocessing out data was congruent to different models. As a result, a lot of the tasks we originally wanted to accomplish got cut out. Thus, for future work, we would like to integrate the SemEval task's original evaluation method of using an hierarchical F1, and we think incorporating visual features into the analysis – multimodal integration – would also be a nice addition that significantly improves performance. Additionally, future work could include data diversification from multiple platforms, as well as refinement in annotations to address the earlier limitations.

6 Contribution Statement

Madeline: data pre-processing, distillbert; wrote models, experiments, related work, and conclusion sections

Izzy: training and evaluation, roberta; wrote introduction, datasets, and results/analysis sections

References

Giovanni Da San-Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles.

Dimitar Dimitrov, Giovanni Da San Martino, Preslov Nakov, Firoj Alam, Maram Hasanain, Abul Hasnat, and Fabrizio Silvestri. 2023. Semeval2024 shared task 4: "multilingual detection of persuasion techniques in memes.".

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

Jakub Piskorski, Nicolas Steganovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. News categorization, framing, and persuasion techniques: Annotation guidelines.

Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka-wei Lee. 2023. Promptmtopic: Unsupervised multimodal topicmodeling of memes using large language models.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper, and lighter.