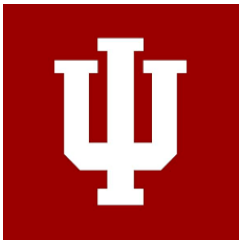

Indiana University Bloomington
Project Report

SemEval: Task 4
Multi-label Hierarchical Classification of Persuasion Techniques in Memes

Submitted by

Priti Singh
Pooja Parab
Siddhant Dilip Godshalwar

Under the guidance of

Dr. Damir Cavar
CSCI-B:659
Advanced Natural Language Processing

1 Abstract

In this research, we delve into the intricate challenge of classifying persuasion techniques within memes, a domain where information is deliberately crafted to fulfill specific agendas. Our primary objective is to discern and categorize these persuasive strategies, encompassing both logical fallacies and emotional appeals, present in the amalgamation of textual and visual elements within memes. The hierarchical structure embedded in the classification system introduces an additional layer of complexity to the task, necessitating a nuanced approach. As memes increasingly serve as influential mediums for communication, understanding the underlying mechanisms of persuasion becomes imperative. By addressing this multifaceted task, we contribute to the broader discourse on digital communication, media literacy, and computational linguistics, shedding light on the intricate interplay between textual and visual elements in the propagation of persuasive content within the unique medium of memes.

Contents

2 Introduction

2.1 Background

This research is conducted within the framework of SemEval, a prominent platform for evaluating and advancing the state-of-the-art in natural language processing (NLP) and computational linguistics. Specifically, the investigation addresses a SemEval 2024 task, focusing on the intricate challenge of classifying persuasion techniques within memes. In the contemporary landscape of digital communication, memes stand as influential vehicles for disseminating information, often shaped by intentional messaging and persuasive tactics.

The SemEval 2024 task revolves around identifying these persuasion techniques embedded in memes, encompassing logical fallacies and emotional appeals present in both textual and visual elements. Embracing a hierarchical classification system, this task adds a layer of complexity, necessitating a sophisticated and nuanced approach to discern the subtleties within meme content accurately.

2.2 Problem Statement

The problem statement for this task involves identifying and categorizing persuasion techniques used in memes, both in their textual and multimodal (textual and visual) content. The task is divided into three subtasks:

Subtask 1: Textual Persuasion Technique Identification
Input: Text extracted from memes.
Output: Identify which of the 20 persuasion techniques, organized hierarchically, are used in the textual content of a meme. Techniques are organized in a hierarchy, and if the ancestor node of a technique is selected, only a partial reward is given. This is a hierarchical multilabel classification problem.

Subtask 2a: Multimodal Persuasion Technique Identification
Input: Text extracted from memes and the image of the memes.
Output: Identify which of the 22 persuasion techniques, organized hierarchically, are used both in the textual and visual content of a meme. Techniques are organized hierarchically, and if the ancestor node of a technique is selected, only partial reward will be given. This is a hierarchical multilabel classification problem.

Subtask 2b: Binary Persuasion Technique Identification in Multimodal Setting
Input: Text extracted from memes and the image of the memes.
Output: Identify whether the meme contains at least one of the 22 persuasion techniques or no technique. This is a binary classification problem, with two classes: "propagandistic" and "non-propagandistic."

Persuasion techniques are methods employed in propaganda, where information is purposefully shaped to achieve a predetermined agenda. These techniques include logical fallacies, emotional language, and other rhetorical and psychological strategies. Memes, which consist of images superimposed with text, are analyzed to understand how these techniques are employed, either in the text or in conjunction with the visual content of the meme.

For this case, this is the hierarchy that we will be trying to implement:

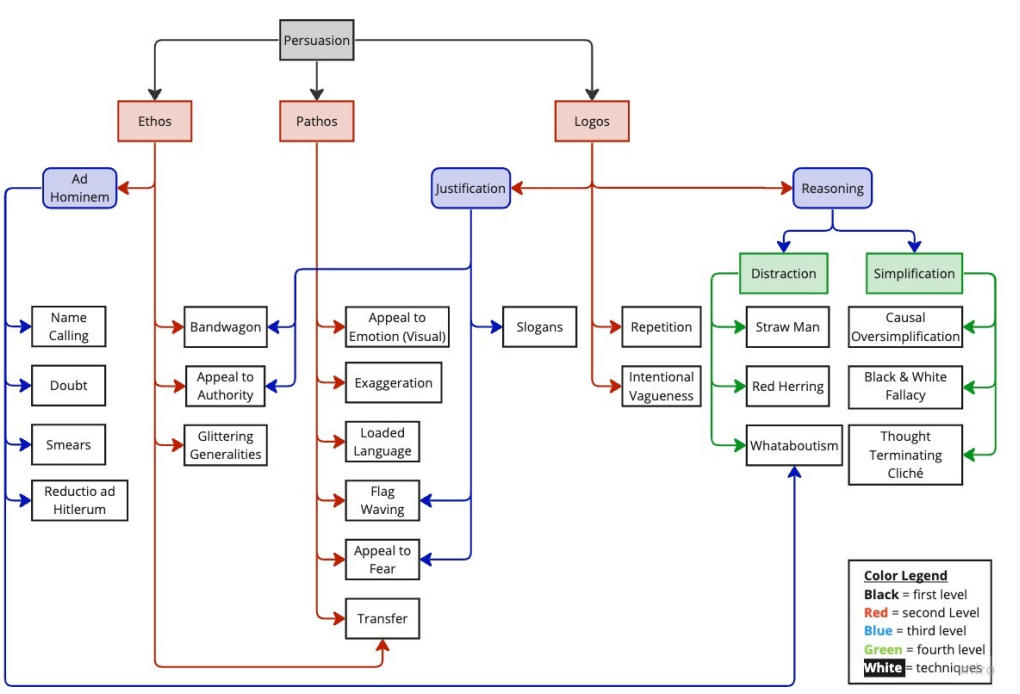

Figure 1: Hierarchy of the techniques for Subtask 2a (in Subtask 1 "Transfer" and ""Appeal to Strong emotion" are not present).

We are also provided with training, development, and test sets in JSON format, containing textual content and, in the case of Subtasks 2a and 2b, images of memes. The goal is to develop models that can accurately identify and classify the persuasion techniques used in memes, considering the hierarchical structure of these techniques. Additionally, surprise test datasets in different languages are introduced to test zero-shot approaches.

For our implementation, we will only be focusing on Subtask 1.

2.3 Objective

In Subtask 1, the objective is to develop models that can accurately identify and categorize persuasion techniques used in the textual content of memes. The input for this subtask is the text extracted from memes, and the goal is to predict which of the 20 persuasion techniques, organized in a hierarchy, are employed in the provided text.

Here are the key components of the objective:

Input:

- The input consists of textual content extracted from memes.
- The text is formatted as a single UTF-8 string, with each sentence on a separate row and blocks of text in different areas of the image separated by a blank row.
- No image data is provided for Subtask 1, making it a natural language processing (NLP) task.

Output:

- The output should identify which persuasion techniques are present in the textual content of a meme.
- The techniques are organized hierarchically, and the goal is to predict the specific technique(s) used.
- The output is a list of technique names and a partial reward is given if the ancestor node of a technique is selected.

Hierarchical Multilabel Classification:

- This is a hierarchical multi-label classification problem where each meme may belong to multiple classes (techniques).
- The hierarchical structure means that techniques are organized in a directed acyclic graph, and the model needs to navigate this hierarchy to make predictions.
- Partial reward is given if a higher-level ancestor of a technique is predicted.

Evaluation:

- Models will be evaluated based on their ability to correctly identify the persuasion techniques in the textual content of memes.
- The evaluation considers both precision and recall, as well as the hierarchical structure of the techniques.

The ultimate objective is to develop models that can automatically analyze textual content in memes and classify the specific persuasion techniques employed, contributing to the identification of propaganda or disinformation strategies in online content.

3 Literature Review

Persuasion, as a fundamental aspect of communication, has been extensively studied throughout history. Aristotle's modes of persuasion—ethos, pathos, and logos—have played a pivotal role in understanding the art of convincing an audience. In addition to these broad categories, scholars have delved into subcategories such as ad hominem, justification, and reasoning, examining how these techniques contribute to persuasive communication.

3.1 Ethos: (2nd Level)

Ethos, characterized by the credibility and authority of the speaker, remains a central focus in the literature on persuasion. Scholars have explored how establishing credibility enhances the effectiveness of persuasive messages. Cialdini (2009) emphasizes the importance of perceived expertise, trustworthiness, and goodwill in building ethos. Research also delves into the impact of non-verbal cues, body language, and ethical considerations on the audience's perception of ethos (Burgoon et al., 2016).

3.2 Pathos: (2nd Level)

The emotional appeal of pathos is a potent force in persuasive communication. Studies have investigated the role of emotions in decision-making and how pathos influences attitudes and behaviors. Slovic et al. (2007) explore the emotional factors that shape risk perception, illustrating the profound impact of emotional appeals on persuasive outcomes. Additionally, scholars have examined the ethical implications of leveraging emotions in persuasion, balancing the line between ethical and manipulative practices (Petty et al., 2018).

3.3 Logos: (2nd Level)

Logos, grounded in logic and reasoning, has been extensively examined in the literature on persuasion. Researchers have explored the cognitive processes underlying logical arguments and how they contribute to attitude change. Tindale (1999) discusses the significance of effective reasoning, identifying fallacies and pitfalls that may undermine logical persuasion. The literature also highlights the role of evidence, statistics, and rational discourse in constructing persuasive messages (Walton, 2006).

3.4 Further Subcategories:

3.4.1 Ad Hominem: (3rd Level)

Ad hominem arguments involve attacking the character of the opponent rather than addressing the substance of their argument. This subcategory of persuasion has garnered attention due to its prevalence in political discourse (Benoit, 1998). Studies have explored the impact of ad hominem attacks on audience perceptions and the ethical implications of using personal attacks as a persuasive strategy (Friggieri, 2019).

3.4.2 Justification: (3rd Level)

Justification as a subcategory of persuasion involves providing reasons and explanations to support a claim. Researchers have examined the role of justification in shaping attitudes and beliefs (Feinberg & Willer, 2011). Additionally, studies have investigated how the quality and relevance of justifications influence the persuasiveness of messages (Allen & Preiss, 1997).

3.4.3 Reasoning: (3rd Level)

Reasoning, closely linked with logos, encompasses the use of sound arguments and logical thinking to persuade an audience. The literature explores various forms of reasoning, including inductive and deductive reasoning, and their impact on persuasive outcomes (Perloff, 2010). Researchers also delve into the role of counterarguments and refutations in strengthening persuasive messages (Moyer-Gusé, 2008).

4 Methodology

4.1 Data Collection:

In conducting our study, we employed a meticulously curated dataset tailored for the meme persuasion technique classification task. This dataset comprises textual content paired with memes, and accompanying labels denote the employed persuasion techniques. The data collection process entailed sourcing memes from diverse online platforms, ensuring a comprehensive representation of content and persuasive strategies. After the collection, the dataset underwent thorough preparation steps to guarantee uniformity and relevance to the nuanced requirements of the classification task. This curated dataset forms the foundation for our investigation, providing a rich and varied source of information essential for training and evaluating classification models.

4.2 Feature Extraction:

Within the feature extraction phase, our emphasis lies on the textual content of memes as the principal feature for classification. The chosen features were carefully derived from the meme text, encompassing linguistic patterns and context crucial for discerning persuasion techniques. To ensure optimal input for our classification model, pre-processing steps were meticulously applied to the data. This encompassed tokenization, breaking the text into meaningful units, and padding, a technique for standardizing sequence lengths. By systematically preparing the data in this manner, we aimed to enhance the model's capacity to capture and analyze the nuanced textual nuances pivotal for effective persuasion technique classification.

4.3 Model Architecture:

4.3.1 Model 1 (BERT):

Our multi-label classification model is built upon the 'owaiskha9654/Multi-Label-Classification-of-PubMed-Articles' BERT model. The architecture includes an input layer for tokenized article text, a BERT layer, and a dense layer for classification with one unit per label. The decision to use this model was influenced by its pre-training on PubMed articles, making it well-suited for the classification of biomedical text.

4.3.2 Model 2 (LSTM):

The model architecture comprises several layers for the multi-label classification task. It starts with an Embedding layer that maps input sequences of words to dense vectors of fixed size, followed by two Bidirectional Long Short-Term Memory (LSTM) layers that capture sequential patterns bidirectionally. To prevent overfitting, Dropout layers are incorporated after each Bidirectional LSTM layer with a dropout rate of 0.5. The final layer is a Dense layer with a sigmoid activation function, representing the output layer for multi-label classification. The model is compiled using the Adam optimizer with a learning rate of 1e-3 and utilizes binary crossentropy as the loss function. The training is performed using a balanced dataset generated by the MultilabelBalancedRandomSampler, and the model undergoes 20 epochs. Evaluation on the validation set involves predicting class probabilities, converting them to binary predictions using a threshold of 0.5, and calculating accuracy as a performance metric.

5 Experiments and Results

5.1 Experimental Setup

5.1.1 Model 1 (BERT):

To assess the efficacy of our model, the data set underwent a meticulous division into training, validation, and testing sets. Essential hyperparameters, including a learning rate of 0.00001 and weight decay of 0.01, were judiciously chosen. The training process spanned 10 epochs, each comprising a batch size of 32. This systematic approach to data set partitioning and parameter selection aimed to optimize model training, facilitating robust evaluation and ensuring the generalization of our multi-label hierarchical classification model across diverse data sets.

5.1.2 Model 2 (LSTM):

The core of the experimental setup lies in the definition and utilization of the MultilabelBalancedRandomSampler class. This custom sampler facilitates the creation of a balanced training dataset by oversampling minority classes and undersampling majority classes, ensuring that each class has at least batch_size / n_classes samples. This novel sampling strategy aims to address imbalances inherent in multi-label data sets.
The TensorFlow model is constructed using an Embedding layer, two Bidirectional LSTM layers, Dropout layers for regularization, and a Dense layer with sigmoid activation for multi-label classification. The model is compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the evaluation metric.
During training, the model is fitted to the balanced training data set for 20 epochs. The performance is evaluated on the validation set by predicting class probabilities, converting them to binary predictions, and calculating accuracy. The entire experimental setup is designed to address challenges posed by imbalanced multi-label data sets, leveraging a customized sampling approach to enhance model robustness and generalization.

5.2 Evaluation Metrics

5.2.1 Model 1 (BERT):

To rigorously evaluate our model, we utilized categorical cross-entropy loss as the optimization objective. This objective function facilitated the fine-tuning of model parameters towards minimizing the discrepancy between predicted and actual labels. Concurrently, we employed categorical accuracy as the evaluation metric, offering insights into the model's proficiency in accurately classifying various persuasion techniques. This dual evaluation approach ensured a comprehensive understanding of the model's performance across the hierarchical spectrum of persuasion techniques within the meme classification task.

5.2.2 Model 2 (LSTM):

The evaluation metrics for the LSTM-based multi-label classification model on the validation set reveal insights into its performance across various classes. The precision, recall, and F1-score metrics are provided for each label, reflecting the model's ability to correctly identify instances of specific rhetorical strategies. The precision metric measures the accuracy of positive predictions, while recall assesses the model's ability to capture all relevant instances. The F1-score provides a balance between precision and recall.

5.3 Results

5.3.1 Model 1 (BERT):

Throughout the training epochs, we meticulously monitored the model's accuracy on both the training and validation sets, elucidating its learning trajectory. The evaluation results, detailed in the classification report, offer nuanced insights into the model's performance across individual persuasion technique labels. A comparative analysis against baseline approaches serves to underscore the efficacy of our hierarchical classification model, providing a robust validation of its capability to discern and classify persuasive techniques within meme content.
Here are some screenshots showing what we have achieved with the model:

```
...  16/16 [==============================] - 5s 132ms/step
     (500, 20)
     Classification Report:
                   precision    recall  f1-score   support

                0       0.19      0.97      0.31        63
                1       0.07      0.89      0.13        27
                2       0.02      0.86      0.04         7
                3       0.14      0.85      0.24        53
                4       0.05      0.86      0.10        21
                5       0.04      0.58      0.08        24
                6       0.06      0.78      0.12        27
                7       0.11      0.86      0.19        42
                8       0.09      0.86      0.17        36
                9       0.34      0.84      0.49       135
               10       0.01      0.75      0.02         4
               11       0.31      0.87      0.45       116
               12       0.00      0.50      0.01         2
               13       0.01      0.75      0.02         4
               14       0.01      1.00      0.02         4
               15       0.06      0.87      0.11        23
               16       0.09      0.60      0.16        50
               17       0.34      0.78      0.47       142
               18       0.09      0.82      0.17        38
               19       0.06      0.90      0.11        21
     ...
        macro avg       0.11      0.81      0.17       839
     weighted avg       0.21      0.82      0.31       839
      samples avg       0.07      0.61      0.12       839
```

Figure 2: Classification Report for Model using BERT

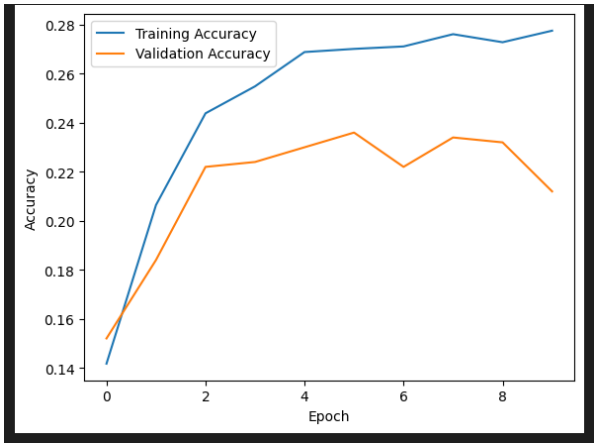

Figure 3: A graph portraying the accuracies over increasing epochs

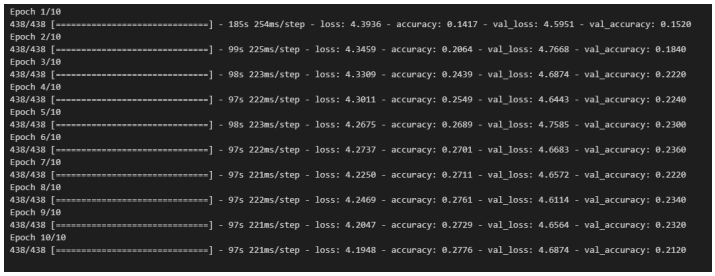

Figure 4: Accuracy of 24% for model using BERT

5.3.2 Model 2 (LSTM):

The model exhibits varying performance across classes, with higher precision, recall, and F1-scores for labels like "Loaded Language" and "Name calling/Labeling." However, challenges arise for labels such as "Appeal to fear/prejudice" and "Causal Oversimplification," reflected in lower metrics. The micro-average precision, recall, and F1-score are 0.43, 0.23, and 0.30, indicating moderate overall performance. The macro-average F1-score is 0.12, highlighting imbalances. Validation accuracy stands at 20.4%, suggesting alignment with ground truth labels in one-fifth of instances. Fine-tuning is recommended for enhanced performance, especially for classes with lower metrics.

Classification Report on Validation Set:

	precision	recall	f1-score	support
Appeal to authority	0.50	0.29	0.36	63
Appeal to fear/prejudice	0.00	0.00	0.00	27
Bandwagon	0.00	0.00	0.00	7
Black-and-white Fallacy/Dictatorship	0.12	0.04	0.06	53
Causal Oversimplification	0.00	0.00	0.00	21
Doubt	0.10	0.04	0.06	24
Exaggeration/Minimisation	0.00	0.00	0.00	27
Flag-waving	0.56	0.12	0.20	42
Glittering generalities (Virtue)	0.33	0.11	0.17	36
Loaded Language	0.45	0.41	0.43	135
Misrepresentation of Someone's Position (Straw Man)	0.00	0.00	0.00	4
Name calling/Labeling	0.50	0.32	0.39	116
Obfuscation, Intentional vagueness, Confusion	0.00	0.00	0.00	2
Presenting Irrelevant Data (Red Herring)	0.00	0.00	0.00	4
Reductio ad hitlerum	0.00	0.00	0.00	4
Repetition	1.00	0.04	0.08	23
Slogans	0.50	0.16	0.24	50
Smears	0.44	0.40	0.42	142
Thought-terminating cliché	0.20	0.05	0.08	38
Whataboutism	0.00	0.00	0.00	21
micro avg	0.43	0.23	0.30	839
macro avg	0.23	0.10	0.12	839
weighted avg	0.37	0.23	0.26	839
samples avg	0.26	0.18	0.20	839

Figure 5: Classification Report for Model using LSTM

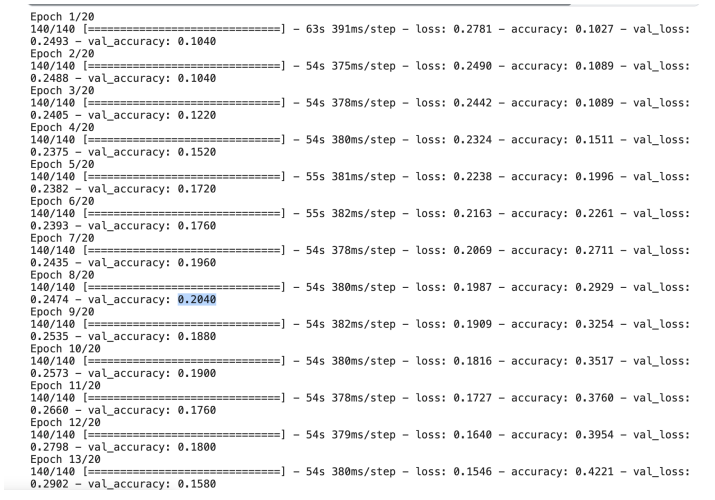

Figure 6: Accuracy of 20.4% for model using LSTM

6 Discussion

6.1 Analysis:

6.1.1 Model 1 (BERT):

The model displays robust recall values across various classes, indicating its capacity to identify a substantial portion of actual positive instances. Notably, classes with higher support, like class 9 and 17, demonstrate strong recall, showcasing the model's effectiveness in capturing prevalent patterns within these categories. The micro-average precision is also reasonable, implying a satisfactory level of accuracy in positive predictions across the entire dataset. Weighted average metrics underscore a balanced performance, considering each class's contribution based on its support. These positive aspects suggest that, with further refinement, the model holds promise for successful classification, particularly in terms of recall and overall predictive accuracy.

6.1.2 Model 2 (LSTM):

The validation accuracy is reported at 20.4%, suggesting that the model's predictions align with the ground truth labels for approximately one-fifth of the instances. The comprehensive evaluation metrics offer a nuanced understanding of the LSTM model's strengths and weaknesses in handling multi-label classification tasks on the given dataset. Fine-tuning and further optimization may be explored to enhance performance, particularly for classes with lower precision, recall, and F1-scores

6.2 Challenges Faced:

6.2.1 Model 1 (BERT):

- The inherent challenge of imbalanced data significantly impacted our study, particularly in the context of multi-label classification. Achieving a balanced representation across various persuasion techniques proved challenging due to the nuanced nature of meme content. Despite concerted efforts, data sampling strategies faced limitations, making it difficult to achieve a uniform distribution. This imbalance adds a layer of complexity to the classification task, demanding careful consideration in the interpretation of results and the generalization of the model to diverse, real-world scenarios.
- During the training phase, our implementation of the BERT-based model encountered early signs of overfitting. In response, extensive experimentation was conducted with hyperparameter tuning to mitigate this challenge. Adjustments to parameters such as learning rate and weight decay were explored to strike a balance between model complexity and generalization. These iterative experiments were crucial in fine-tuning the model's performance, ensuring robustness and preventing premature overfitting, ultimately enhancing the model's capacity to effectively classify persuasion techniques within memes.

6.2.2 Model 2 (LSTM):

The LSTM model faces challenges because it has difficulty learning from less common examples, and the dataset is not big enough for it to understand things well. The way it reads and processes words might lose some important details, and it's not easy to figure out why it makes specific predictions. Some categories, like "Loaded Language" and "Name calling/Labeling," are predicted better than others. To make the model better, we need to work on these issues, like handling imbalanced data and choosing better settings for how it reads and understands the text.

7 Conclusion

In conclusion, our investigation into the classification of persuasion techniques within memes has provided valuable insights into the challenges and potential advancements in this domain. Leveraging two distinct models, namely BERT and LSTM, our study addressed the nuanced task of hierarchical multi-label classification, focusing on textual content analysis. Despite encountering challenges, such as imbalanced data and model-specific intricacies, our models exhibited promising strengths. Model 1, based on BERT, showcased robust recall values, especially for classes with higher support, emphasizing its efficacy in capturing prevalent patterns. Model 2, employing LSTM, revealed room for improvement, with a validation accuracy of 20.4%. Notably, both models demonstrated varying performance across specific persuasion technique labels, indicating the need for nuanced strategies in handling imbalances.

The comprehensive evaluation metrics provided a detailed understanding of each model's strengths and weaknesses. Model 1 demonstrated a balanced performance, particularly in terms of recall, while Model 2 exhibited challenges in achieving higher precision and recall across all classes. The exploration of hyperparameter tuning and data preprocessing strategies was instrumental in enhancing model robustness, and addressing issues such as overfitting and imbalanced data.

Challenges faced during the study underscore the complexity of meme classification, highlighting the need for further refinement in handling imbalances and improving model interpretability. Future work should focus on developing more advanced models, exploring diverse pre-trained architectures, and incorporating innovative techniques to address imbalances within the dataset. The ultimate goal remains the development of models that can effectively analyze textual content in memes, contributing to the identification of persuasion techniques and promoting a nuanced understanding of propaganda and disinformation strategies in online content.

8 Future Work

In our pursuit of achieving superior accuracy, we are committed to a multifaceted strategy. Firstly, fine-tuning hyperparameters remains a key focus, involving meticulous adjustments to parameters such as learning rates, weight decay, and batch sizes. Concurrently, we explore variations in model architecture, considering potential enhancements or alternative pre-trained models that could better capture the nuanced features of meme content.

Recognizing the challenges posed by class imbalance, we are dedicated to refining our data preprocessing techniques. This includes exploring advanced methodologies to address and mitigate the imbalance, ensuring a more equitable representation of persuasion techniques in the training data set.

Furthermore, our commitment extends to the implementation of hierarchical classification. This entails a structured approach that leverages the inherent hierarchy within the persuasion techniques, facilitating a more nuanced understanding of the relationships between different classes. By refining these aspects collectively, we aim to elevate the accuracy of our classification model, ultimately advancing its efficacy in discerning persuasion techniques within the complex and dynamic landscape of memes.

9 References

- *Mendoza, S. N. (2023, December 14).* MemePersuasionDetection. . Github
- Hugging Face. (2023). Transformers Documentation: BERT. Retrieved December 14, 2023, from BERT
- PyTorch. (2023). torch.nn.LSTM. Retrieved December 14, 2023, from LSTM
- *Allen, M., & Preiss, R. (1997).* Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. Communication Research Reports, 14(2), 125-131.
- *Benoit, W. L. (1998). William L. Benoit replies.* Political Communication, 15(4), 471-473.
- *Cialdini, R. B. (2009).* Influence: Science and practice. Pearson Education.
- *Feinberg, M., & Willer, R. (2011).* The moral roots of environmental attitudes. Psychological Science, 22(3), 324-328.
- *Friggieri, J. (2019).* The Ad Hominem Fallacy. Philosophy and Rhetoric, 52(4), 388-414.
- *Moyer-Gusé, E. (2008).* Toward a theory of entertainment persuasion: Explaining the persuasive effects of entertainment-education messages. Communication Theory, 18(3), 407-425.
- *Petty, R. E., Briñol, P., & Tormala, Z. L. (2018).* Thought confidence as a determinant of persuasion: The self-validation hypothesis. In The Oxford Handbook of Social Influence (pp. 205-224). Oxford University Press.
- *Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007).* The affect heuristic. European Journal of Operational Research, 177(3), 1333-1352.
- *Tindale, C. W. (1999).* Acts of argument: A contribution to the study of argumentation. Walter de Gruyter.
- *Walton, D. (2006).* Fundamentals of critical argumentation. Cambridge University Press.