Detecting Persuasion Techniques in Memes

Kota Shamanth Ramanath Nayak

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science at

Concordia University

Montréal, Québec, Canada

November 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Kota Shamanth Ramanath Nayak**

Entitled: **Detecting Persuasion Techniques in Memes**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. René Witte

_____ Examiner
Dr. René Witte

_____ Examiner
Dr. Essam Mansour

_____ Supervisor
Dr. Leila Kosseim

Approved by _____
Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2024

_____
Mourad Debabbi, Dean
Faculty of Engineering and Computer Science

Abstract

Detecting Persuasion Techniques in Memes

Kota Shamanth Ramanath Nayak

Memes, which are user-generated content in the form of images and text, have become a powerful medium for shaping public discourse. Given their increasing influence, detecting persuasive techniques embedded within these multimodal forms of communication is crucial for identifying propaganda and combating online disinformation. Persuasion techniques in memes often combine rhetorical elements from both text and image, creating unique challenges for computational models.

This thesis seeks to determine the impact of multimodal integration on the detection of persuasion techniques in memes and to evaluate how well multimodal models perform compared to single-modality models in this classification task. To achieve this, we developed and fine-tuned several models for text-based and multimodal persuasion detection using both pre-trained language models (BERT, XLM-RoBERTa, mBERT) and image-based models (CLIP, ResNET, VisualBERT).

A key contribution of this work is the implementation of paraphrase-based data augmentation, which helped address class imbalance and improved the performance of text-only models. For multimodal approaches, we explored both early fusion and cross-modal alignment strategies. Surprisingly, cross-modal alignment underperformed, likely due to challenges in aligning abstract textual and visual cues. In contrast, the early fusion approach of combining text and image embeddings showed the highest performance, significantly outperforming text-only and image-only models.

We also conducted zero-shot experiments with GPT-4 to benchmark its effectiveness in multimodal persuasion detection. Although GPT-4 demonstrated potential in zero-shot settings, the fine-tuned models still outperformed it, particularly when leveraging multimodal integration.

This research advances the understanding of multimodal learning for detecting persuasion techniques, with broader implications for disinformation detection in online content.

Acknowledgments

I would like to begin by expressing my heartfelt gratitude to Dr. Leila Kosseim, my supervisor, for giving me the opportunity to join this program and for her steadfast support and guidance throughout my journey. Her patience, insight, and endless feedback across the numerous iterations of my thesis and our conference submissions, including SemEval at NAACL 2024, NLPA 2024 and Canadian AI 2024, were instrumental in helping me grow both as a researcher and as a person. Thank you, Leila, for always pushing me to achieve my best.

I am also incredibly thankful to my fellow members of the CLaC Lab: Alejandra Zambrano, Andrei Neagu, Deokyeong Kim, Eeham Khan, Jennifer Marks, Owen Van Esbroeck, Nelson Filipe Costa, and Nawar Turk, for their continuous encouragement and invaluable feedback during all our discussions and dry runs. Special thanks to past lab member Farhood Farahnak, whose advice was greatly appreciated during the initial phases of my work. The supportive environment within our lab made this journey all the more enriching.

To my parents, Ramanath and Shubha, your constant love and belief in me kept me going, and I am eternally grateful for everything you have done. I also want to extend my heartfelt thanks to my friends Akash and Adivardhan for their companionship, laughter, and encouragement, which made all the difference during this journey.

I am deeply grateful to Dr. Rene Witte and Dr. Essam Mansour for their time and effort in reviewing my thesis, as well as for their insightful feedback and thought-provoking questions during my defense.

And, of course, to all my lovely friends—thank you for being there every step of the way, offering support, sending love, and making this experience unforgettable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Memes are user-generated, shareable digital content in the form of images or text, often humorous or satirical, which spread rapidly across social media platforms and contribute to public discourse (Shifman & Handloff, 2015). According to Wikipedia[1], characteristics of memes include their susceptibility to parody, their viral propagation and their evolution over time. Because of this, in today's digital era, memes have emerged as a powerful medium for communication, often blending humor, social commentary, and political discourse. They are a pivotal element in disinformation campaigns. However, their ability to convey persuasive messages, sometimes as part of disinformation campaigns, has raised significant concerns (Milner & Stephens, 2018). Memes, which combine textual and visual elements, present unique challenges for Natural Language Processing (NLP) researchers, particularly to detect the persuasive techniques embedded within these multimodal contents.

Memes have become an influential tool for shaping opinions in online spaces, often leveraging rhetorical and visual techniques to influence users' perceptions. For example, during the 2016 U.S. presidential election, Wendling (2018) found that memes played a significant role in spreading political content, with certain memes on Facebook receiving over 1 million shares, illustrating their powerful reach in influencing public opinion. By analyzing the persuasive elements in memes, we can better understand the strategies used to spread disinformation, bias, and propaganda. The multimodal nature of memes complicates the task of detecting these techniques, as the visual and

[1] https://en.wikipedia.org/wiki/Internet_meme

textual elements often work together to enhance the persuasive message.

The meme in Figure 1.1 exemplifies how text and images can be combined to deploy multiple persuasion techniques. The phrase "TRAITOR JOE'S" utilizes a *Smears* technique (as it attacks an individual's reputation (Joe Biden's) with negative connotations), a *Name calling/Labeling* technique (through the use of the term "traitor", aimed at provoking negative emotions) and a *Transfer* technique (by associating the political figure (Joe Biden) with the flags of countries like Russia and China, creating negative associations that evoke emotional responses from viewers). The combination of these persuasion techniques, along with the use of visual elements like flags that evoke strong political associations, strengthens the persuasive power of the meme.

Figure 1.1: Meme displaying the text *"TRAITOR JOE'S"*, annotated with the persuasion techniques *Smears*, *Transfer*, and *Name calling/Labeling*. (Image source: `prop_meme_18261.png` from the SemEval 2024 Task 4 training dataset)

Previous works such as (Feng et al., 2021) have explored multimodal approaches to the detection of persuasion detection in memes, often considering the interaction between text and image. However, to our knowledge, there has been little investigation into how different types of interactions between these modalities—such as early fusion versus cross-modal approach—specifically

affect the detection of persuasive techniques. This thesis addresses this gap by examining how these distinct interaction strategies impact the effectiveness of persuasion detection, particularly focusing on the unique contributions of textual elements when integrated with or separated from visual components.

1.1 Goal of the Thesis

Over the past few years, the field of Natural Language Processing (NLP) has made significant strides in text classification (Shu, Sliva, Wang, Tang, & Liu, 2017), multimodal learning (L. H. Li, Yatskar, Yin, Hsieh, & Chang, 2019), and disinformation detection (Howard & Ruder, 2018). With the growing importance of memes in shaping online discourse, understanding the persuasive strategies used in these multimodal elements has become increasingly relevant. Traditional NLP techniques have often focused on the textual content only (contained either in the text or in the image directly), while ignoring the synergistic effects that arise when both (text and image) modalities are combined.

This thesis seeks to determine the impact of multimodal integration on the detection of persuasion techniques in memes and to evaluate how well multimodal models perform compared to single-modality models in this classification task.

To achieve this goal, we have developed and compared different approaches to detecting persuasion techniques in memes and measured the contribution of each modality (text and image) to the final decision as part of the recent SemEval 2024 Task 4 challenge (Dimitrov et al., 2024). We have developed models based on fine-tuning various textual and visual pre-trained models—such as BERT, XLM-Roberta, CLIP, and VisualBERT and have evaluated them in a multilabel classification setting using both text and visual meme content.

As the dataset of SemEval 2024 was imbalanced and multilingual, we have experimented with the use of paraphrase-generation data augmentation technique, and explored the model's ability to generalize to unseen languages.

This research contributes to the growing body of work in multimodal learning (Tan & Bansal,

2019), disinformation tracking (S. Yu, Martino, Mohtarami, Glass, & Nakov, 2021) and the detection of persuasion techniques (Da San Martino, 2019) with broader implications for tasks such as sentiment analysis, emotion detection, and content moderation in multimodal environments.

1.2 Methodology

The methodology employed in this thesis involves two key components. First, we developed models by fine-tuning pre-trained language models such as BERT (see Section 2.2.5), XLM-Roberta (see Section 2.2.8), and mBERT (see Section 2.2.6) to detect persuasion techniques in the textual content of memes (see Chapter 3). These models were evaluated on the SemEval 2024 Task 4 hierarchical multilabel classification tasks (Dimitrov et al., 2024), where the goal is to identify 0 or more persuasion techniques from a predefined set of 20 techniques. To address the data imbalance, we employed data augmentation techniques, based on paraphrasing, to increase the representation of underrepresented persuasion techniques.

We then extended the scope of our work to multimodal classification by combining textual and visual inputs (see Chapter 4). Here, we used a variety of architectures, including CLIP (see Section 2.2.11) and VisualBERT (see Section 2.2.9), to explore early fusion and cross-modal alignment strategies. The aim was to determine how well these models perform compared to text-only models. All of the code used in the implementation of the models described in this thesis is made available on GitHub. [2]

1.3 Contributions

This thesis makes several contributions to the field of NLP, particularly in the areas of persuasion detection, multimodal learning, and disinformation tracking:

(1) **Multilabel Classification and Data Augmentation for Persuasion Detection in Text:** We propose and evaluate a multilabel classification system that fine-tuned pre-trained language

[2] https://github.com/CLaC-Lab/SemEval-2024-Task-4

models on textual meme content to detect persuasion techniques based on the dataset proposed in Dimitrov et al. (2024). To address dataset imbalance, we introduce paraphrase-based data augmentation techniques, which resulted in a significant performance boost. Specifically, models trained on the augmented dataset showed improvement in performance compared to those trained on the original, imbalanced dataset. This contribution is described in Chapter 3 and has been published in Nayak and Kosseim (2024a, 2024b).

(2) **Multimodal Persuasion Detection:** We explore the effectiveness of combining textual and visual modalities in persuasion detection, utilizing early fusion strategy and cross-modal alignment. Our multimodal approach, which combines text and image embeddings, surprisingly outperformed both text-only and image-only models, as well as the cross-modal alignment technique. This highlights the importance of capturing the interactions between modalities for more accurate persuasion detection. Details on this work are provided in Chapter 4.

(3) **Zero-shot Experiments with GPT-4:** To better situate the performance of the models, with the emerging capabilities of today's large language models, we evaluated the performance of zero-shot learning through GPT-4 on persuasion detection task, comparing it against our fine-tuned model. While GPT-4 showed potential in zero-shot settings, the fine-tuned model consistently outperformed it in this specialized task, particularly when multimodal interactions were involved. This comparison is discussed in Chapter 4.

Overall, this research contributes to the broader effort to detect and combat online disinformation, particularly in the context of multimodal content like memes, where persuasion techniques are often subtle and complex.

1.4 Thesis Structure

This chapter has outlined the motivation, goals, and contributions of this thesis. Chapter 2 provides an overview of related work in persuasion technique detection, NLP for disinformation tracking, and multimodal and multilabel classification. Chapter 3 details our approach to text-only

multilabel classification of persuasion techniques in meme texts using the SemEval 2024 dataset (Dimitrov et al., 2024). Chapter 4 extends this work to multimodal content, combining text and image data to improve classification performance. Finally, Chapter 5 summarizes the findings of this thesis, its limitations, and proposes potential avenues for future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter provides a review of the key models and methodologies relevant to the detection of persuasion techniques and propaganda.

Section 2.2 examines the architectures of various models across three domains: text models, vision models, and vision-language models. Each of these categories contributes to the foundations upon which more complex and integrated systems for classification and detection are built. Following this, the focus shifts to multi-label classification, exploring the challenges and advancements in handling multiple labels within a dataset. Section 2.4 includes an overview of previous works that have contributed to this field, extending into the realms of multimodal multi-label classification and hierarchical multi-label classification, where intricate relationships between data types and labels are managed. Next, Section 2.5 provides an in-depth look at hierarchical evaluation metrics.

The chapter the provides a detailed discussion of prior research on the detection of propaganda and persuasion techniques. Section 2.6 underscores the growing importance of understanding how these techniques are deployed across different media, and the evolution of models and methods designed to detect and classify such techniques in various contexts

2.2 Architectures

This section presents an overview of key advancements in neural network architectures for natural language processing (NLP), focusing on Convolutional Neural Networks (CNNs) (O'Shea & Nash, 2015), Residual Networks (ResNets) (He, Zhang, Ren, & Sun, 2015), and Transformers (Vaswani et al., 2017), and extending to state-of-the-art models such as BERT (Devlin, Chang, Lee, & Toutanova, 2019), mBERT (Devlin et al., 2019), RoBERTa (Y. Liu et al., 2019), XLM-Roberta (Conneau et al., 2020), VisualBERT (L. H. Li et al., 2019), Vision Transformers (ViTs) (Dosovitskiy et al., 2021), and CLIP (Radford et al., 2021). Understanding these models is necessary to appreciate the rest of the thesis

2.2.1 Feed Forward Neural Networks (FFNNs)

Neural networks (Schmidhuber, 2015) are computational models inspired by the biological neural networks in the human brain. They consist of interconnected nodes (neurons) organized into layers: an input layer, one or more hidden layers, and an output layer. Each connection has an associated weight, which adjusts during training to minimize the difference between the predicted and actual outputs.

A basic feed forward neural network (FFNN) performs the following operations:

- **Forward Pass**: Computes the output by applying weights and activation functions to the input data.
- **Loss Calculation**: Measures the difference between the predicted output and the actual target using a loss function.
- **Backward Pass**: Updates weights through backpropagation based on the loss gradient.

The activation function is crucial in introducing non-linearity, enabling the network to learn complex patterns. Common activation functions include ReLU (Rectified Linear Unit) (Agarap, 2019), Tanh and Sigmoid.

2.2.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (O'Shea & Nash, 2015) are a specialized type of neural network designed for processing structured grid data. They have been used extensively for image processing but also in NLP (Jacovi, Sar Shalom, & Goldberg, 2018; Kim, 2014). They use convolutional layers to detect local patterns and features within the data. Key components of CNNs include:

- **Convolutional Layers**: Apply convolutional filters to the input data to extract features. The convolution operation is defined as:

$$(I * K)(x, y) = \sum_m \sum_n I(x + m, y + n) \cdot K(m, n)$$

  where I is the input (an image or a sentence), K is the convolutional kernel, and (x, y) are the coordinates of the output feature map.

- **Activation Layers**: Apply non-linear activation functions like ReLU to introduce non-linearity.

- **Pooling Layers**: Reduce the spatial dimensions of the feature maps, typically using operations like max pooling or average pooling.

- **Fully Connected Layers**: Flatten the feature maps and apply a dense layer made of a FFNN (see Section 2.2.1) to produce the final classification output.

CNNs have become the backbone of image recognition tasks due to their ability to capture hierarchical features but have also been used in NLP (Jacovi et al., 2018; Kim, 2014) with less wide-spread use.

2.2.3 Residual Networks (ResNets)

Residual Networks, introduced by He et al. (2015), address the challenge of training very deep neural networks by incorporating residual connections. These connections allow gradients to flow more easily through the network, facilitating the training of deeper models.

The core idea of ResNets is the residual block, which includes shortcut connections that bypass one or more layers. The residual function is given by:

$$y = \mathcal{F}(x, \{W_i\}) + x$$

where:

- $\mathcal{F}(x, \{W_i\})$ represents the residual function learned by the block,
- x is the input to the block.

By learning identity mappings, ResNets improve gradient flow and mitigate the vanishing gradient problem, allowing for the effective training of very deep networks.

2.2.4 Transformers: A Paradigm Shift

Transformers, introduced Vaswani et al. (2017), revolutionize sequence modeling by using self-attention mechanisms instead of recurrent or convolutional layers. This allows transformers to capture long-range dependencies and complex relationships within sequences.

The self-attention mechanism computes the attention scores for each token relative to others in the sequence:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where:

- Q is the query matrix,
- K is the key matrix,
- V is the value matrix, and
- d_k is the dimensionality of the key vectors.

Multi-head attention enables the model to focus on different parts of the input simultaneously, enhancing its ability to capture diverse patterns. Positional encoding is added to input embeddings to retain the order of tokens, which is crucial for understanding the sequence context.

2.2.5 BERT: Bidirectional Encoder Representations from Transformers

BERT (Bidirectional Encoder Representations from Transformers), (Devlin et al., 2019), improves contextual understanding by processing sequences bidirectionally. Unlike previous models (eg: the vanilla transformer of Section 2.2.4) that processed text in a unidirectional manner, BERT considers both left and right context.

BERT's pretraining involves two main tasks:

- **Masked Language Modeling (MLM)**: Some tokens in the input sequence are masked, and the model learns to predict these masked tokens. The loss function for MLM is:

$$L_{MLM} = -\sum_{i \in M} \log P(x_i \mid x_{/i})$$

  where:

  - x_i is a masked token,
  - $x_{/i}$ denotes the remaining tokens (i.e., the context around x_i), and
  - M is the set of indices of masked tokens.

- **Next Sentence Prediction (NSP)**: The model predicts whether a given sentence follows another, aiding in understanding sentence relationships.

$$L_{NSP} = -[y \log P(\text{is\_next}) + (1 - y) \log P(\text{not\_next})]$$

  where:

  - y is the true label (1 if the second sentence follows the first, 0 otherwise),
  - $P(\text{is\_next})$ is the predicted probability that the second sentence follows the first, and
  - $P(\text{not\_next})$ is the predicted probability that the second sentence does not follow the first.

  This task helps BERT better understand the relationship between sentences, improving its ability to model longer text sequences.

BERT's bidirectional approach enables it to create contextual embeddings which can be used as a basis to build models with strong performance across various NLP tasks by capturing richer contextual information.

2.2.6 mBERT: Multilingual BERT

mBERT (Multilingual BERT) (Devlin et al., 2019) extends BERT's capabilities to multiple languages by training on a diverse corpus that includes 104 languages. This multilingual model leverages shared linguistic features, enabling it to perform well across different languages.

mBERT's multilingual training allows it to handle cross-lingual tasks effectively and exhibit zero-shot learning capabilities, making it robust in multilingual settings.

2.2.7 RoBERTa: Robustly Optimized BERT Pretraining Approach

RoBERTa (Robustly Optimized BERT Pretraining Approach), (Y. Liu et al., 2019), builds upon BERT with several improvements:

- **Larger Training Data**: RoBERTa is trained on a larger dataset (160GB of text) compared to BERT, including data from BooksCorpus and English Wikipedia (16GB).
- **Longer Training Duration**: RoBERTa benefits from longer training with larger batch sizes.
- **Dynamic Masking**: Unlike BERT's static masking, RoBERTa uses dynamic masking, where the masked tokens change in each epoch.
- **Removal of NSP**: RoBERTa removes the Next Sentence Prediction (NSP) objective used in BERT, finding that it does not significantly impact performance.

These improvements lead to better language understanding and representation capabilities, contributing to superior performance on various benchmarks.

2.2.8 XLM-Roberta: A Robustly Optimized BERT Pretraining Approach for Multilingual Data

XLM-Roberta (Cross-lingual Language Model - RoBERTa) (Conneau et al., 2020) extends RoBERTa to handle multilingual data. It builds on RoBERTa's architecture and improvements but is trained on a more extensive and diverse corpus, including 2.5 TB of text from 100 languages.

XLM-Roberta incorporates:

- **Multilingual Training**: Utilizes a corpus that spans 100 languages, enhancing cross-lingual performance.
- **SentencePiece Tokenization**: Employs SentencePiece tokenization (Kudo & Richardson, 2018) to handle diverse linguistic structures.
- **Shared Encoder**: Uses a single encoder for all languages, learning common representations across different languages.

These enhancements enable XLM-Roberta to achieve state-of-the-art performance in multilingual settings and improve cross-lingual understanding.

2.2.9 VisualBERT: Visual and Language Representation Learning

VisualBERT, introduced by L. H. Li et al. (2019), integrates visual and textual information by extending BERT to handle both modalities. It processes images and text together, capturing interactions between visual and textual features. VisualBERT features are:

- **Unified Encoder**: Combines image features (extracted using a visual backbone like Faster R-CNN (Ren, He, Girshick, & Sun, 2016)) with textual embeddings.
- **Cross-Modal Attention**: Aligns and integrates visual and textual features to understand their relationships. The cross-attention mechanism is defined as:

$$\text{Cross-Attention}(Q_{\text{image}}, K_{\text{text}}, V_{\text{text}})$$

where:

- Q_{image} is the query matrix from image features,
- K_{text} and V_{text} are the key and value matrices from text embeddings.

VisualBERT's ability to process and align visual and textual data makes it effective for multimodal tasks such as image captioning and visual question answering.

2.2.10 Vision Transformers: Transforming Computer Vision

Vision Transformers (ViTs), introduced by Dosovitskiy et al. (2021), adapt the transformer architecture for image processing. ViTs treat image patches as tokens and apply self-attention mechanisms to capture global dependencies within images. The Vision Transformer process includes:

- **Image Patching**: Dividing an image into fixed-size patches.
- **Token Embedding**: Linearly embedding these patches into a sequence of tokens.
- **Transformer Encoder**: Processing the sequence of tokens through a transformer encoder.
- **Class Token**: Adding a learnable class token to the sequence for classification purposes.

ViTs leverage the self-attention mechanism to capture global contextual information in images, achieving state-of-the-art performance in various computer vision tasks.

2.2.11 CLIP: Contrastive Language–Image Pretraining

CLIP (Contrastive Language–Image Pretraining), developed by Radford et al. (2021), aligns images and textual descriptions using contrastive learning.[1] It employs separate encoders for images and text, training them to map into a shared embedding space.

The image encoder, based on a vision transformer (see Section 2.2.10), converts images into embeddings, while the text encoder processes textual descriptions with a transformer model (see Section 2.2.4). The contrastive loss function used is:

$$L = -\log \frac{\exp(\text{sim}(I, T_{pos})/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(I, T_i)/\tau)}$$

[1] Contrastive learning emphasizes the extraction of representations from data by highlighting the differences between positive (similar) and negative (dissimilar) pairs of instances.

where:

- $\text{sim}(I, T)$ represents the similarity score between image I and text T,
- T_{pos} denotes the positive (matching) text,
- T_i represents the negative (non-matching) texts, and
- τ is the temperature parameter.

CLIP's approach enables strong zero-shot learning capabilities, allowing the model to generalize across various image and text tasks without task-specific training.

2.3 Advances in Large Language Models: GPT Series

The development of large language models (LLMs) has significantly advanced NLP tasks. The Generative Pre-trained Transformer (GPT) series, developed by OpenAI, exemplifies these advancements by leveraging transformer-based architectures (Vaswani et al., 2017) and pretraining on extensive datasets. This section provides an overview of the GPT series, detailing its architectural innovations and capabilities, culminating in GPT-4's advancements.

2.3.1 GPT: The Foundation

The original GPT (Radford, 2018) introduced a unidirectional transformer architecture for text generation. Its decoder-only design facilitated autoregressive text generation by attending only to previous tokens in a sequence. Key architectural features included:

- **Transformer Blocks**: Each block comprised multi-head self-attention layers and feedforward networks.
- **Positional Encoding**: Positional embeddings accounted for token order within the input sequence.
- **Pretraining-Finetuning Paradigm**: The model was pretrained on large-scale textual data and fine-tuned on specific tasks, demonstrating transfer learning's potential in NLP.

Although GPT's capacity (110 million parameters) limited its performance on complex tasks, it laid the foundation for scaling LLMs.

2.3.2 GPT-2: Scaling Up

GPT-2 (Radford et al., 2019) expanded upon GPT by increasing model size (1.5 billion parameters) and training on a larger, diverse dataset. Its innovations included:

- **Deeper Networks**: The model had 48 transformer layers in its largest configuration.
- **Larger Context Windows**: The ability to process sequences of up to 1024 tokens improved coherence over extended texts.
- **Improved Zero-Shot Learning**: Without task-specific fine-tuning, GPT-2 demonstrated the ability to perform tasks such as text generation and summarization by interpreting prompts.

GPT-2's capabilities marked a significant improvement, achieving state-of-the-art performance on various NLP benchmarks.

2.3.3 GPT-3: Massive Scaling and In-Context Learning

GPT-3 (Brown et al., 2020) scaled the model size to 175 billion parameters, leveraging increased computational power and dataset diversity. Architectural innovations included:

- **Sparse Attention Optimizations**: Efficient self-attention mechanisms allowed scaling without prohibitive computational costs.
- **Layer Normalization**: Improved stability during training by normalizing representations within transformer layers.
- **In-Context Learning**: GPT-3 could perform tasks by conditioning on a few input examples (zero-shot or few-shot), eliminating the need for fine-tuning.

These enhancements allowed GPT-3 to perform a wide range of NLP tasks, including text classification, summarization, and question answering, with remarkable generalization abilities.

2.3.4 GPT-3.5: Refinement and Efficiency

GPT-3.5 built upon GPT-3's architecture, focusing on training efficiency and task-specific coherence. Although OpenAI has not released detailed architecture-specific papers for GPT-3.5, improvements included:

- **Mixed Precision Training**: Leveraging 16-bit floating-point precision improved training speed and memory efficiency.
- **Fine-Tuned Adaptation**: Enhanced fine-tuning techniques allowed more effective downstream task performance.

A key innovation introduced during this phase was **Reinforcement Learning from Human Feedback (RLHF)**. RLHF aligns the model's behavior with human expectations by leveraging human feedback during fine-tuning. This three-step process involves supervised fine-tuning on labeled datasets, training a reward model based on human feedback, and reinforcement learning to optimize the model's outputs. RLHF significantly improved GPT-3.5's ability to generate contextually appropriate, human-aligned responses, laying the foundation for further advancements in GPT-4.

2.3.5 GPT-4: Enhanced Zero-Shot Learning and Applications

GPT-4, the latest model in the GPT series, exemplifies the advances in large-scale LLMs by further increasing model capacity, improving generalization, and expanding its range of applications (OpenAI et al., 2024). Its innovations include:

- **Increased Model Capacity**: GPT-4 builds upon its predecessors by incorporating a significantly larger number of parameters.
- **Extensive Pretraining**: Training on diverse and expansive datasets enabled the model to generate highly coherent and contextually relevant text.
- **Enhanced Zero-Shot Learning**: GPT-4 demonstrates exceptional performance across tasks without explicit fine-tuning by leveraging its broad general knowledge base.

An essential aspect of GPT-4's success is the integration of **Reinforcement Learning from Human Feedback (RLHF)**, which ensures that the model's outputs align with human values and preferences. By refining the reward model and optimizing through reinforcement learning, RLHF enhanced GPT-4's ability to generate safe, coherent, and contextually appropriate responses. This innovation has been particularly impactful in tasks requiring nuanced reasoning, ethical considerations, and human-like understanding.

2.3.6 GPT-4 Vision: Multimodal Learning

GPT-4 Vision extends GPT-4's architecture to support multimodal tasks, enabling joint reasoning over text and visual data. Its architectural enhancements include:

- **Visual Tokenization**: Images are tokenized into patch embeddings or feature maps compatible with transformer processing.
- **Cross-Modality Attention**: Specialized attention layers facilitate interactions between text and visual features, allowing coherent multimodal reasoning.
- **Unified Architecture**: Visual and textual encoders produce aligned embeddings, enabling tasks like visual question answering and image-based reasoning.

By integrating visual capabilities, GPT-4 Vision has broadened the scope of LLMs, demonstrating effectiveness in multimodal tasks such as image captioning and visual content summarization.

2.3.7 Prompt Engineering and its Evolution

Prompt engineering has emerged as a crucial aspect of working with large language models. It involves designing and optimizing input prompts to guide LLMs towards generating accurate, contextually appropriate outputs for specific tasks. The quality of prompts can significantly influence the performance of LLMs in both zero-shot and few-shot settings. Early work by Radford et al. (2019) introduced the idea of using prompts to frame tasks for LLMs, and this approach has since evolved, with GPT-3 and GPT-4 models showcasing how task performance can be dramatically improved through well-designed prompts.

Researchers have explored various strategies for prompt engineering, including manual prompt crafting and automated approaches such as prompt tuning (Lester, Al-Rfou, & Constant, 2021). Manual crafting involves using human intuition to design task-specific prompts that elicit the desired response from the model. This method is particularly effective for zero-shot learning, where task-specific data may be unavailable, and the model must rely on the prompt's wording to infer the task. In contrast, prompt tuning is an automated approach that fine-tunes the prompt parameters to achieve optimal task performance. Studies have shown that prompt tuning can rival or surpass traditional fine-tuning techniques in specific tasks, making it a valuable tool in leveraging pre-trained models like GPT-4 (P. Liu et al., 2021).

2.3.8 Chain-of-Thought Reasoning

A notable advancement in prompting techniques is the introduction of Chain-of-Thought (CoT) reasoning, which allows large language models to perform complex reasoning tasks by generating intermediate reasoning steps rather than providing a direct answer (Wei et al., 2023). CoT prompting encourages the model to break down a problem into smaller, more manageable components, leading to more accurate and explainable outputs.

Studies show that CoT prompting improves performance on tasks that require multiple inference steps, such as arithmetic problem-solving or commonsense reasoning (Nye et al., 2021). By asking the model to "think" through the problem explicitly, CoT helps to mitigate common errors associated with direct question answering and enhances the interpretability of the model's reasoning process. Recent work has also demonstrated that CoT reasoning can be combined with self-consistency techniques, where multiple reasoning paths are generated, and the model's output is selected based on the most frequent or plausible answer among the generated paths (Wang et al., 2023).

2.4 Multi-label Classification

Multi-label classification is a type of machine learning problem where each instance may belong to multiple classes simultaneously, as opposed to traditional single-label classification where each instance is assigned to only one class. Multi-label classification is particularly useful in various

applications where multiple "attributes" [2] are relevant, such as in text categorization, image tagging, and medical diagnosis.

In multi-label classification, the goal is to predict a set of labels for each instance from a predefined set of possible labels. The output for each instance is typically represented as a binary vector where each element corresponds to a label and indicates its presence (1) or absence (0). For instance, in a multi-label text classification task with labels such as 'sports', 'politics', and 'technology', an article about both 'sports' and 'technology' would be represented by the vector $[1, 0, 1]$.

2.4.1 Previous Works in Multi-label Classification

Several methods and approaches have been developed for tackling multi-label classification, especially in text data. These can be divided into non-neural and neural approaches.

Non-Neural Approaches:

- **Binary Relevance**: This approach treats each label as a separate binary classification problem. Each label is predicted independently of the others. While simple and easy to implement, this method does not capture the dependencies between labels. K. Tsoumakas and Katakis (2007) discussed the binary relevance approach and highlighted its simplicity and limitations in their comprehensive overview.

- **Classifier Chains**: This method extends binary relevance by modeling dependencies between labels. A chain of binary classifiers is used, where each classifier in the chain predicts a label based on the input features and the predictions of previous classifiers. Read, Pfahringer, Holmes, and Frank (2009) demonstrated that classifier chains could improve performance by capturing label dependencies, although the approach may suffer from error propagation.

- **Label Powerset**: Label powerset treats each unique set of labels as a separate class in a multi-class classification problem. This approach can capture complex label correlations but becomes computationally expensive as the number of possible label combinations increases.

[2] In this context, the word "attributes" refers to the different characteristics or categories that an instance can be associated with.

G. Tsoumakas, Katakis, and Vlahavas (2010) introduced this method and discussed its applicability and limitations.

Neural Approaches:

Neural networks, including Convolutional Neural Networks and Recurrent Neural Networks, are commonly adapted for multi-label classification tasks. In these models, the output layer often uses a sigmoid activation function to predict the probability of each label independently.

The Binary Cross-Entropy loss function is typically used for multi-label classification tasks. This function calculates the loss for each label separately and averages it across all labels. The loss function is defined as:

$$Loss = -\frac{1}{L}\sum_{i=1}^{L}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{1}$$

where:

- y_i is the binary indicator (0 or 1) for label i,
- p_i is the predicted probability for label i,
- L is the total number of labels.

Several neural approaches have been proposed:

Deep Neural Networks: Deep learning models such as Convolutional Neural Networks and Recurrent Neural Networks have been employed for multi-label classification tasks. Q. Li et al. (2022) proposed using deep learning architectures for text classification and demonstrated significant improvements over traditional methods. Their approach leverages the ability of deep models to capture complex patterns in the data.

Transformer-based Models: Transformers, particularly BERT and its variants, have been adapted for multi-label classification. BERT, provides a powerful framework for capturing contextual information and has been modified for multi-label tasks by adjusting the output layer and using appropriate loss functions. Cheng et al. (2021) further extended this approach by applying

transformer-based models to multi-label classification, showing that pre-trained language models can enhance performance.

Attention Mechanisms: Attention mechanisms allow models to focus on relevant parts of the input, which can be beneficial for multi-label classification tasks. Yang et al. (2016) incorporated attention mechanisms into their multi-label classification model, improving the handling of label correlations and enhancing overall performance.

Several studies have explored multi-label classification of textual content. In 2019 (Chalkidis, Fergadiotis, Malakasiotis, & Androutsopoulos, 2019) showed that Bi-GRUs with label-wise attention led to good performance, and the inclusion of domain-specific Word2vec and context-sensitive ELMo embeddings further boosted the performance on the EURLEX57K dataset that contained 57k English EU legislative documents. Lin, Qin, Wang, Zhou, and Yang (2023) introduced five innovative contrastive losses for multi-label text classification using the dataset from the SemEval 2018 Multi-label Emotion Classification (MEC) task (Mohammad, Bravo-Marquez, Salameh, & Kiritchenko, 2018) in English, Arabic, and Spanish that contained 8640 instances. All five contrastive learning methods notably enhanced the performance of the previous top-performing model, SpanEmO (Alhuzali & Ananiadou, 2021) for the MEC task.

2.4.2 Multimodal Research in Multi-label Classification

Multimodal research, i.e combining different types of data (e.g., text and images), has been explored in multi-label classification performance. While leveraging multiple modalities can provide richer representations and enhance predictions, success is not guaranteed. The effectiveness of this approach depends on the quality and relevance of each modality, and in some cases, the additional data might introduce noise or complexity that can hinder performance. Therefore, multimodal methods have the potential to improve outcomes, but their impact varies based on the specific context and task. Three main approaches have been proposed:

- **Early Fusion**: This approach combines features from different modalities before classification. Features extracted from images and text can be concatenated or merged into a unified

representation, which is then passed through a classifier. This method allows for the integration of diverse information but requires careful feature engineering and alignment.

For example: in image captioning using a multimodal model, features from an image may be extracted using a ResNet model (see Section 2.2.3), while features from the corresponding text can be extracted using a BERT model (see Section 2.2.5). These feature vectors are concatenated and passed through a FFNN (see Section 2.2.1) for classification or regression tasks, such as predicting the correct caption. This approach ensures that both visual and textual features are fused early and processed together through a unified model.

- **Late Fusion**: In this approach, separate models are trained for each modality, and their predictions are combined at a later stage. Pooling techniques such as voting or averaging are used to aggregate predictions from different modalities. Late fusion can effectively leverage specialized models for each modality but may not capture complex interactions between modalities.

  For example: in multimodal sentiment analysis, separate models can be used to analyze text and images. For instance, a BERT model can be used to analyze the sentiment in the text, while a VGG model (Simonyan & Zisserman, 2015) processes the image. At the final stage, the predictions (probabilities or labels) from both models are combined using a voting mechanism or an averaging method to produce a final sentiment prediction.

- **Cross-Modal Attention**: Cross-modal attention mechanisms align and integrate features from different modalities, allowing the model to capture interactions between them. This approach enhances the model's understanding of multimodal data by focusing on relevant features across modalities.

  For example: in VisualBERT (see Section 2.2.9), visual features can be extracted from images using Faster R-CNN (Ren et al., 2016), while text features can be processed by BERT. Cross-modal attention layers align the visual and textual features, enabling the model to integrate both modalities effectively. This approach allows the model to focus on relevant parts of the image in relation to the text, improving its performance in multimodal tasks like visual question answering or image captioning.

2.4.3 Hierarchical Multi-label Classification

In hierarchical multi-label classification (HMC), samples are assigned to one or more class labels within a structured hierarchy (see Figure 2.1). Approaches to HMC can be divided into local and global methods. Local methods use multiple classifiers, often overlooking the overall structure of the hierarchy. For example, Cerri, Barros, and de Carvalho (2014) trained a multi-layer perceptron incrementally for each hierarchy level, using predictions from one level as inputs for the next. In contrast, global methods employ a single model to address all classes and implement various strategies to capture the hierarchical relationships between labels. One such approach, Zhou et al. (2020), modeled the hierarchy as a directed graph and introduced hierarchy-aware structure encoders, using a bidirectional TreeLSTM and a hierarchy-GCN to extract and aggregate label structural information in an end-to-end fashion. C. Yu, Shen, Mao, and Cai (2022) redefined hierarchical text classification (HTC) as a sequence generation task and developed a sequence-to-tree (Seq2Tree) framework to model the hierarchical label structure. Additionally, they created a constrained decoding strategy with a dynamic vocabulary to ensure label consistency.

2.5 Hierarchical Evaluation Measures

In this thesis, we address the multi-label classification of persuasion techniques from the shared task SemEval 2024 Task 4 (Dimitrov et al., 2024). This task employed hierarchical evaluation measures which we will now explain through an example.

In this example, we consider a simplified classification hierarchy for animals. The hierarchy includes both flying and flightless birds, flying mammals like bats, and reptiles. The full structure is shown in Figure 2.1.

To account for cases where animals can share characteristics with multiple groups, we introduce a directed acyclic graph (DAG) structure. For example, the *Bat*, while classified as a *mammal*, also shares the characteristic of flight with birds. This makes *Bat* a child of both the *Flying Mammals* and *Flying Birds* categories.

To evaluate multi-label hierarchical classifiers, we calculate hierarchical precision, recall, and F1 score, which take into account the ancestor nodes of both the ground truth and predicted labels.

Below is an example scenario where we compute these metrics.

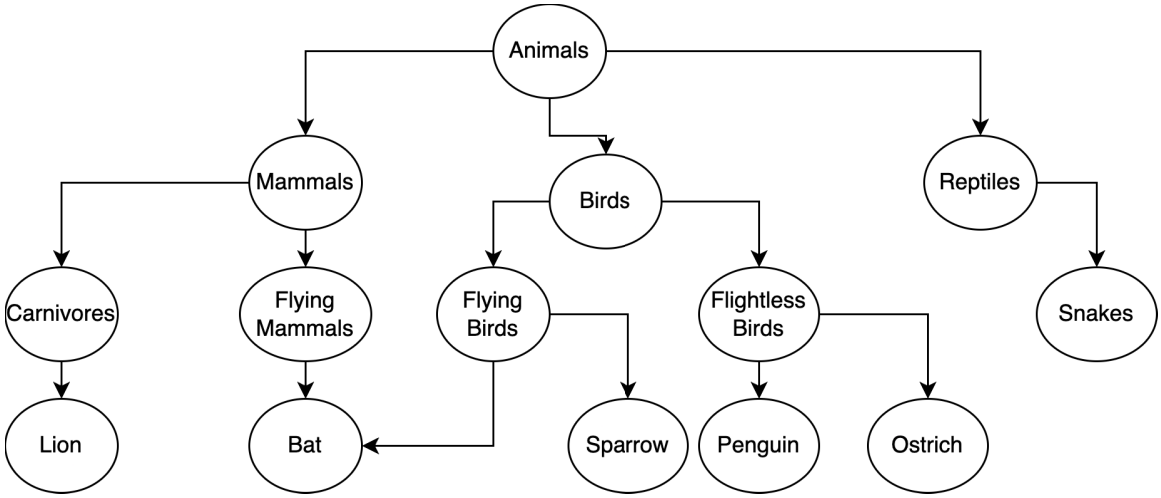

Figure 2.1: Example graph for hierarchical evaluation of Multi-Label Classification

Scenario:

- Ground Truth (G): *Bat*

- Prediction (P): *Penguin*

For the ground truth label "Bat", the set of ancestors (S_{gold}) includes:

$$S_{gold} = \{\text{Bat}, \text{Flying Mammals}, \text{Mammals}, \text{Flying Birds}, \text{Birds}, \text{Animals}\}$$

For the predicted label "Penguin", the set of ancestors (S_{pred}) includes:

$$S_{pred} = \{\text{Penguin}, \text{Flightless Birds}, \text{Birds}, \text{Animals}\}$$

Hierarchical Precision:

Hierarchical precision (hP) is calculated as the intersection of the ancestors of the predicted label and the ground truth label, divided by the total number of ancestors of the predicted label:

$$hP = \frac{|S_{gold} \cap S_{pred}|}{|S_{pred}|} = \frac{|\text{Birds}, \text{Animals}|}{|\text{Penguin}, \text{Flightless Birds}, \text{Birds}, \text{Animals}|} = \frac{2}{4} = \frac{1}{2}$$

Hierarchical Recall:

Hierarchical recall (hR) is calculated as the intersection of the ancestors of the predicted label and the ground truth label, divided by the total number of ancestors of the ground truth label:

$$hR = \frac{|S_{gold} \cap S_{pred}|}{|S_{gold}|} = \frac{|\text{Birds, Animals}|}{|\text{Bat, Flying Mammals, Mammals, Flying Birds, Birds, Animals}|} = \frac{2}{6}$$

Hierarchical F1 Score:

The hierarchical F1 score (hF) is calculated as the harmonic mean of hierarchical precision and hierarchical recall:

$$hF = \frac{2 \cdot hP \cdot hR}{hP + hR} = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{6}}{\frac{1}{2} + \frac{2}{6}} = \frac{\frac{4}{12}}{\frac{5}{6}} = \frac{2}{5}$$

These metrics have been used as part of the SemEval 2024 Task 4 (Dimitrov et al., 2024) and will be used to evaluate our results in Chapters 3 and 4.

2.6 Propaganda and Persuasion Techniques

Persuasion techniques play a crucial role in communication, often being employed to spread propaganda. Due to their significance and today's ease at generating text via Large Language Models (LLMs), there has been a growing body of research within NLP aimed at detecting these techniques. This has led to significant progress in the identification and analysis of propaganda through automated methods.

Persuasion techniques have become increasingly prominent in digital propaganda campaigns and as a consequence several inventories of techniques have been developed. For example, the inventory from the European Commission (Piskorski, N. Stefanovitch, et al., 2023), the short inventory from Goffredo, Chaves, Villata, and Cabrio (2023) and the more commonly used one from Dimitrov et al. (2024). Da San Martino (2019) emphasizes that modern propaganda extends beyond fake news, encompassing various rhetorical techniques like logical fallacies and emotional appeals. This broader perspective on propaganda detection has led to the development of methods that address the complexity of these persuasive tactics. However, despite advancements, one of the main challenges is the scarcity of high-quality annotated datasets, particularly at the fragment level (i.e., labeling specific text segments like phrases or clauses rather than entire sentences), which impedes

the straightforward application of deep learning models in this area.

2.7 Developments in Propaganda Detection through Shared Tasks

Most recent work in propaganda detection has been done within the context of shared tasks, which have provided structured datasets and benchmarks to compare different approaches. These tasks have evolved over time, from identifying propaganda in complete documents to more complex challenges like fine-grained detection at the fragment level and real-time detection systems.

2.7.1 NLP4IF 2019 and Early Progress in Propaganda Detection

The NLP4IF 2019 Shared Task (Da San Martino, Barrón-Cedeño, & Nakov, 2019) laid the foundation for propaganda detection by introducing two subtasks:

- Fragment-Level Classification (FLC): Detecting specific propaganda techniques within text fragments.
- Sentence-Level Classification (SLC): Classifying whether entire sentences contained propaganda.

These tasks attracted a wide range of approaches, from traditional machine learning models to neural models, laying the groundwork for future developments. The top-performing teams employed BERT-based models for both FLC and SLC tasks. Team Newspeak (Yoosuf & Yang, 2019) achieved the highest performance in the FLC task using a 20-way word-level classification with BERT. They experimented with unsupervised fine-tuning on news datasets and oversampling techniques, reaching an F1 score of 0.25. Meanwhile, Team LTUorp (Mapes, White, Medury, & Dua, 2019) excelled in the SLC task with an attention-based BERT model trained on Wikipedia and BookCorpus, achieving an F1 score of 0.63.

2.7.2 SemEval-2019 and Hyperpartisan News Detection

Following the success of NLP4IF 2019, the SemEval-2019 Task 4 (Kiesel et al., 2019) focused on detecting hyperpartisan news, a form of media manipulation that aligns closely with propaganda.

This task aimed to identify extreme political biases in news articles, challenging systems to detect manipulative rhetoric and biased content. Team Bertha von Suttner (Jiang, Petrak, Song, Bontcheva, & Maynard, 2019) achieved an F1 score of 0.81 using ELMo embeddings combined with a convolutional neural network (CNN). Their system was designed to preprocess minimal text and used a combination of dense layers and ensembles for final predictions.

This shared task helped pave the way for further research by highlighting the challenges in identifying biased and hyperpartisan content, which is often closely related to propaganda.

2.7.3 SemEval 2020 and the Shift Toward Fine-Grained Detection

In SemEval 2020 Task 11 (Da San Martino, Barrón-Cedeño, Wachsmuth, Petrov, & Nakov, 2020), the focus shifted toward more fine-grained detection of propaganda. Two subtasks were introduced:

- Span Identification (SI): Identifying specific propagandistic fragments within news articles.
- Multi-Label Technique Classification (TC): Classifying multiple propaganda techniques within these fragments.

Team Hitachi (Morishita, Morio, Ozaki, & Miyoshi, 2020) led the span identification task with an F1 score of 0.52 using BIO encoding and an ensemble of pre-trained language models. Team ApplicaAI (Jurkiewicz, Borchmann, Kosmala, & Graliński, 2020) achieved the best results in the technique classification task (F1 of 0.62) using a RoBERTa-CRF architecture with self-supervised learning. The fine-grained approach in this shared task marked a significant evolution from earlier tasks, focusing not just on detecting propaganda but on identifying the exact techniques used within fragments.

2.8 Advances in Real-Time and Fine-Grained Propaganda Detection

Building on these shared tasks, researchers began to explore real-time and fine-grained detection systems. This allowed for more precise detection of specific techniques and real-time intervention to counter disinformation.

2.8.1 Proppy: A Real-Time Detection System

The Proppy system (Barrón-Cedeño, Da San Martino, Jaradat, & Nakov, 2019), built on the success of earlier systems that identified hyperpartisan news and propaganda, introduced a real-time approach to detecting propaganda in online news. Unlike previous systems that focused primarily on analyzing static datasets, Proppy continuously monitors online news outlets, clusters articles around specific events, and ranks them based on their likelihood of containing propaganda.

Proppy utilizes word n-grams, lexical and stylistic markers, and vocabulary richness to compute a "propaganda index" for each article. This system achieved state-of-the-art performance in distinguishing propagandistic content and provided a practical application for organizing and evaluating news articles in real time. The approach aligns with previous systems, such as those used in SemEval-2019, but innovates by offering real-time insight into propagandistic intent.

2.8.2 Fine-Grained Propaganda Detection and Explainability

One of the most important developments following Proppy was the shift toward fine-grained propaganda detection, which took root in (Da San Martino, Yu, Barrón-Cedeño, Petrov, & Nakov, 2019). Unlike prior efforts that labeled entire articles or news outlets as propagandistic or not, this approach broke down content into a set of 18 constituent techniques, such as *Loaded language*, *Name calling/labelling*, and *Appeals to fear/prejudice*.

The key innovation in this work was the introduction of a large annotated corpus and a multi-granularity neural network, which outperformed BERT-based baselines. By identifying specific propaganda techniques within fragments, this approach provided more nuanced evaluations and enhanced explainability in AI-driven detection systems.

2.9 Recent Trends in Propaganda Detection

2.9.1 Evolving Techniques and Fragment-Level Detection

Further advancements in fine-grained detection were made by S. Yu, Martino, and Nakov (2019), who focused on fragment-level detection. Their research built upon systems like Proppy and expanded the analysis to 18 distinct propaganda techniques, enabling models to perform more granular analysis of propagandistic content.

By designing a multi-granularity neural network that leveraged sentence-level and fragment-level information, this work enhanced both the precision and recall of propaganda detection. This system represented a significant leap forward from previous baseline models, integrating different layers of granularity to identify propaganda techniques more effectively.

2.9.2 Multimodal Propaganda Detection in the COVID-19 Era

The COVID-19 pandemic saw a significant rise in misinformation and disinformation, often referred to as the *infodemic*. During this time, Nakov and Da San Martino (2021) explored how propaganda detection systems could be adapted to combat fake news, media bias, and misinformation during the pandemic. Their work emphasized the importance of developing tools to detect multimodal propaganda, including memes and other harmful content.

Building on earlier developments in propaganda detection, the study also highlighted the challenges posed by increasingly sophisticated disinformation campaigns, such as those powered by neural fake news generators like GPT-3 (Brown et al., 2020). This necessitated a move toward more comprehensive detection strategies that included fact-checking, stance detection, and source reliability estimation.

2.10 Advancing Explainability and Multilingual Detection

2.10.1 Interpretable Propaganda Detection

As the field matured, the focus shifted toward making propaganda detection systems more interpretable. S. Yu et al. (2021) introduced a framework for enhancing the interpretability of these

models. Their approach combined syntactic and semantic information to provide better explanations of why certain text fragments were labeled as propagandistic.

By integrating human behavior insights, such as sentence positioning and topic similarity, this work advanced both the performance and explainability of propaganda detection models, helping to build trust in automated systems.

2.10.2 SemEval-2023: Multilingual Propaganda Detection

In SemEval-2023 Task 3 (Piskorski, Stefanovitch, Da San Martino, & Nakov, 2023), researchers took on the challenge of detecting persuasion techniques, the framing, and categorizing the genre of online news articles in a multilingual setup. This task included articles from nine languages and focused on global topics like the COVID-19 pandemic and the Russo-Ukrainian war.

The task introduced three subtasks: categorizing news genres, detecting framing from a set of 14 dimensions, and identifying persuasion techniques across 23 specific categories. Among the top teams, Wu et al. (2023) excelled across multiple languages by using an ensemble of fine-tuned mBERT models and task-adaptive MLM pretraining. This team performed especially well in handling surprise languages like Georgian. Their success was aided by the use of additional satire resources and carefully engineered ensembles.

Another notable team, MarsEclipse (Liao, Lai, & Nakov, 2023), implemented a multi-label contrastive loss fine-tuning strategy on XLM-RoBERTa. Their approach excelled in languages like Italian and German by adapting the loss function to a multi-label setup, allowing them to effectively handle both the framing and persuasion techniques tasks.

This multilingual approach represents a significant step forward in the realm of propaganda detection, showing that it is not only essential to detect propaganda in a single language but also to address the complexity of detecting it across multiple languages and cultural contexts.

As discussed earlier, most research on detecting persuasion techniques has primarily focused on a single modality, specifically text. However, in 2024, a new shared task (SemEval 2024 Task 4 (Dimitrov et al., 2024)) introduced a multimodal approach, incorporating both text and image. This thesis aims to explore this dual-modality question.

2.11 Chapter Summary

This chapter provided a review of foundational models, methodologies, and key advancements in the detection of persuasion techniques and propaganda, focusing on text, vision, and multimodal systems. It began by examining key architectures, including neural networks, CNNs, ResNets, Transformers, and advanced models like BERT, RoBERTa, and XLM-Roberta. These models transformed NLP, vision, and multimodal tasks, laying the foundation for modern classification systems.

The chapter then discussed multi-label classification, covering traditional approaches like binary relevance and classifier chains, as well as advanced neural models designed to handle multiple labels simultaneously. Special attention was given to hierarchical multi-label classification, where relationships between labels were critical, especially for detecting persuasion techniques.

In terms of multimodal classification, the chapter highlighted the integration of text and image data using early fusion, late fusion, and cross-modal attention mechanisms. This was particularly relevant to detecting persuasion techniques in complex media like memes, where combining modalities enhanced performance.

The review proceeded with an examination of developments in propaganda detection through shared tasks, starting with NLP4IF 2019 (Da San Martino, Barrón-Cedeño, & Nakov, 2019) and SemEval-2019 (Kiesel et al., 2019), which focused on fragment and sentence-level detection of propaganda, introducing advanced BERT-based approaches. SemEval 2020 (Da San Martino et al., 2020) marked a shift toward fine-grained detection, where teams used models like RoBERTa and CRFs to classify specific propaganda techniques at the fragment level.

Building on these tasks, the chapter covered the Proppy system, which introduced real-time monitoring of online news for propaganda detection. This system innovated by continuously tracking articles and ranking them for propagandistic content, addressing the growing challenge of disinformation.

The chapter concluded by exploring recent trends, including multilingual and multimodal detection. SemEval-2023 (Piskorski, Stefanovitch, et al., 2023) was highlighted for its multilingual setup, where teams tackled persuasion techniques across nine languages, showcasing advancements

in cross-lingual detection and contrastive learning. The chapter emphasized the increasing complexity of propaganda detection in digital and multilingual environments, highlighting the importance of explainability and real-time capabilities.

The next chapter will focus on the methodology for detecting persuasion techniques in meme texts, with a particular emphasis on Subtask 1 of SemEval 2024 Task 4 (Dimitrov et al., 2024).

Chapter 3

Multi-label Classification of Persuasion Techniques in Meme Texts

3.1 Introduction

This chapter provides an in-depth description of our methodology for identifying persuasion techniques within meme texts. The findings presented here were published at the SemEval Workshop co-located with the NAACL conference (Nayak & Kosseim, 2024b) and the 5th International Conference on Natural Language Processing and Applications (NLPA) (Nayak & Kosseim, 2024a).

Section 3.2 outlines the task objectives. Section 3.3 presents the datasets used in this study, highlighting the various sources of meme texts and their annotations for persuasion techniques. In Section 3.4, we describe our proposed approach, which involves fine-tuning pre-trained language models on augmented datasets to detect persuasion techniques in meme texts. Our methodology includes threshold-based multi-label classification to assign zero or more persuasion techniques to each text, along with custom thresholds for each technique to optimize detection. We also implemented data augmentation strategies to address class imbalance, improving model performance across underrepresented techniques. Following this, Section 3.5 outlines the experimental setup, including the architecture of our models and the hyperparameters that were fine-tuned during training. Finally, in Section 3.6, we present the results of our experiments, offering an analysis of the model's performance and insights into the effectiveness of our methodology.

3.2 The SemEval 2024 Task 4

The SemEval-2024 shared Task 4 (Dimitrov et al., 2024) introduced three distinct subtasks aimed at uncovering how memes use various persuasion techniques to shape user perspectives. Subtask 1 focused solely on the analysis of textual content, while Subtasks 2 and 3 examined the multimodal context, integrating both textual and visual elements. Subtasks 1 and 2 employed hierarchical multi-label classification metrics, whereas Subtask 3 involved a binary classification task. Although the training dataset provided was in English, all subtasks required evaluating our model's zero-shot performance on three surprise languages (which turned out to be Bulgarian, North Macedonian, and Arabic) as well as an additional dataset in English. The testing phase aimed to assess the model's ability to generalize to these languages without explicit training. This chapter specifically details our participation in Subtask 1, which focuses on the detection of 20 hierarchically structured persuasion techniques within the textual content of memes.

3.3 Datasets

The goal of Subtask 1 was to categorize only the textual content of memes into zero or several persuasion techniques. An inventory of 20 techniques was provided (eg: *Smears*, *Loaded Language*, *Slogans*) and were structured hierarchically, rendering the task a hierarchical multi-label classification problem. For example, given the training instance shown in Figure 3.1, the model needs to learn that the text *Don't expect a broken government to fix itself* should be labelled with the three techniques provided in the labels field.

```
{
    "id": "79369",
    "text": "Don't expect a broken government to fix itself.",
    "labels": [
        "Loaded Language",
        "Slogans",
        "Name calling/Labeling"
    ],
    "link": "null"
}
```


Figure 3.1: A sample training instance for SemEval 2024 Task 4. The text is labelled with three techniques, *Loaded Language*, *Slogans* and *Name calling/Labelling*.

The SemEval organizers collected memes in English, Bulgarian, North Macedonian, and Arabic from their personal Facebook accounts, scraping public groups discussing politics, vaccines, COVID-19, gender equality, and the Russo-Ukrainian War. For Subtask 1, the input data included only the text extracted from these memes. The training (7k samples), validation (500 samples) and development (1k samples) sets included only English texts; whereas the test set was multilingual with 1500 samples for English, 426 samples for Bulgarian, 259 samples for North Macedonian and 100 samples for Arabic. All datasets were the provided in the form of JSON files. The dataset was annotated with the 20 persuasion techniques shown in Figure 3.2. As the figure shows, the techniques (the leaf nodes) are organised in a graph. Appendix A lists the definitions of these persuasion techniques.

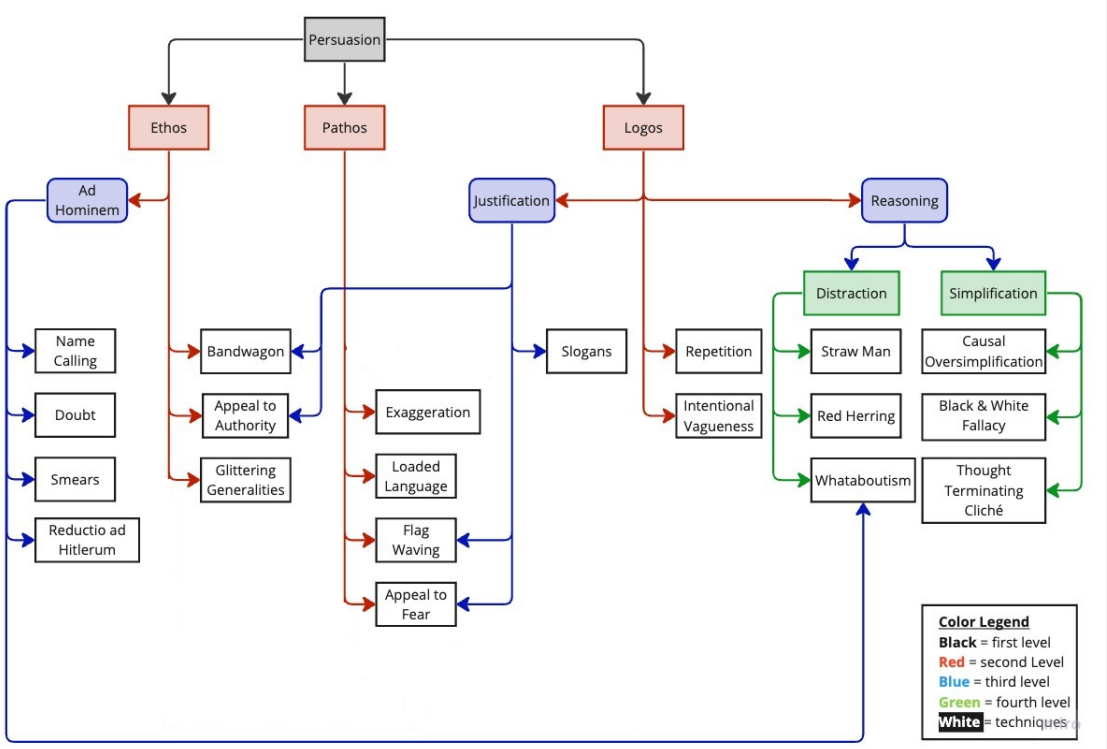

Figure 3.2: Graph of the persuasion techniques to be used to label the texts. Figure slightly modified from (Dimitrov et al., 2024).

Figure 3.3 shows the distribution of the data for each persuasion technique in the training set. As the figure shows, some techniques, such as *Loaded Language* and *Smears*, had a substantial number of samples, while others like *Straw Man* and *Red Herring* were severely underrepresented.

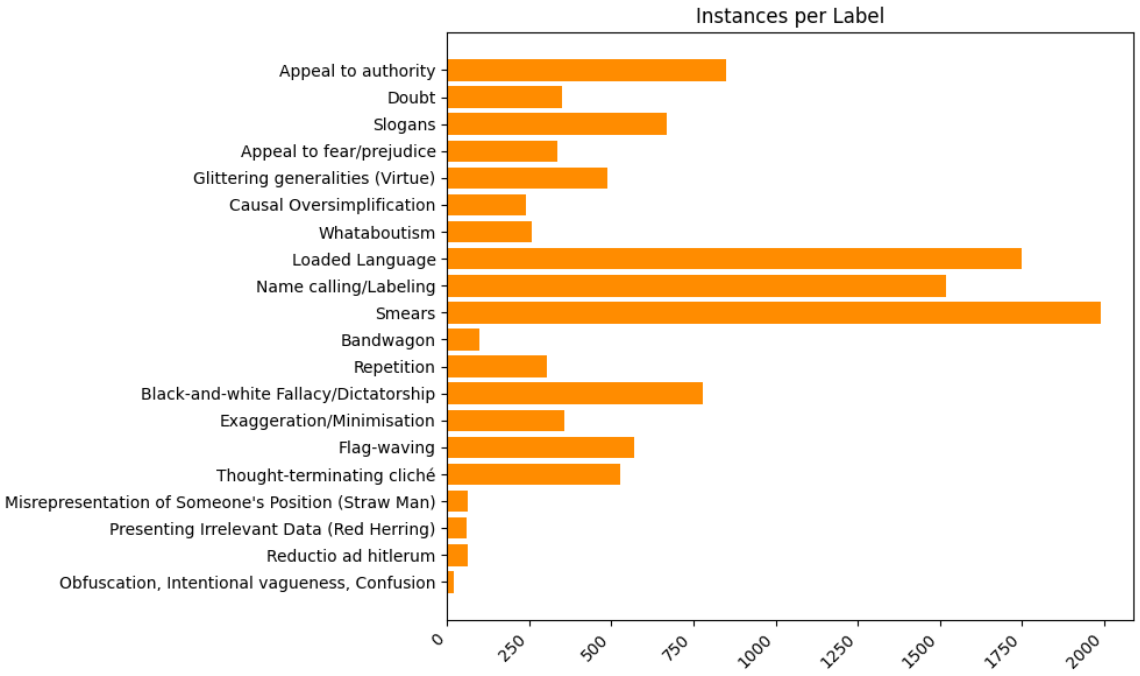

Figure 3.3: Distribution of the data for each persuasion technique in the SemEval 2024 training set.

Figure 3.4 shows the distribution of labels per instances. As the figure shows, most of the instances (47%) were labelled with multiple techniques, 35% were labeled with only 1 technique and 18% had no labels at all.

Given the above English training set and hierarchical persuasion techniques as shown in Figure 3.2, the goal of our model was to identify 0 or n techniques for each textual instance in English and in three surprise languages.

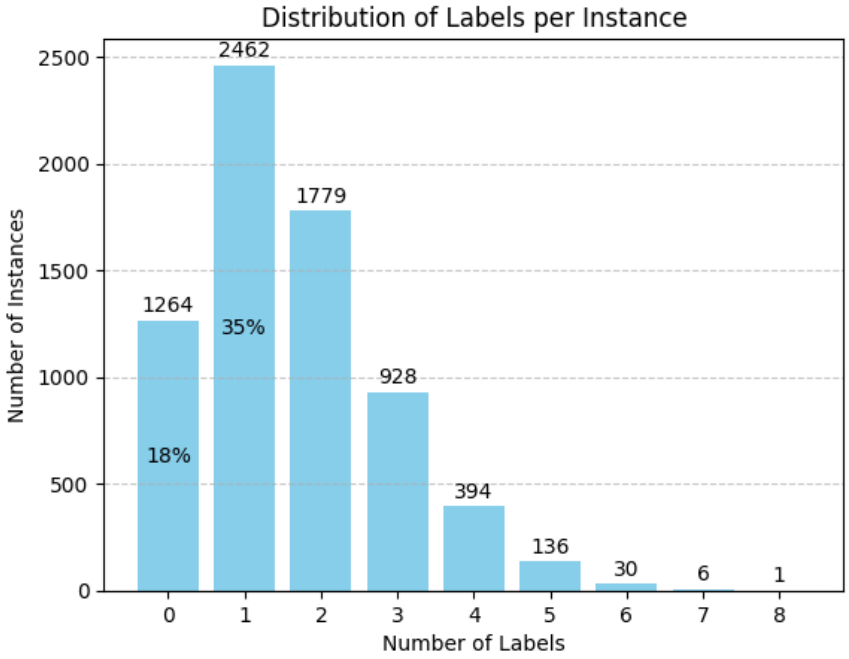

Figure 3.4: Distribution of persuasion techniques per instance in the training set.

3.4 Proposed Approach

Figure 3.5 shows an overview of the classification pipeline we employed for this subtask. As shown in the Figure 3.5, our methodology is based on fine-tuning three distinct pre-trained language models: BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and mBERT (Devlin et al., 2019). This fine-tuning process is conducted on augmented datasets. The data is first preprocessed using the WordPiece Tokenizer. Then we proceeded to fine-tune the three distinct models which returned a probability distribution over the 20 techniques. These three model predictions were then pooled via averaging.

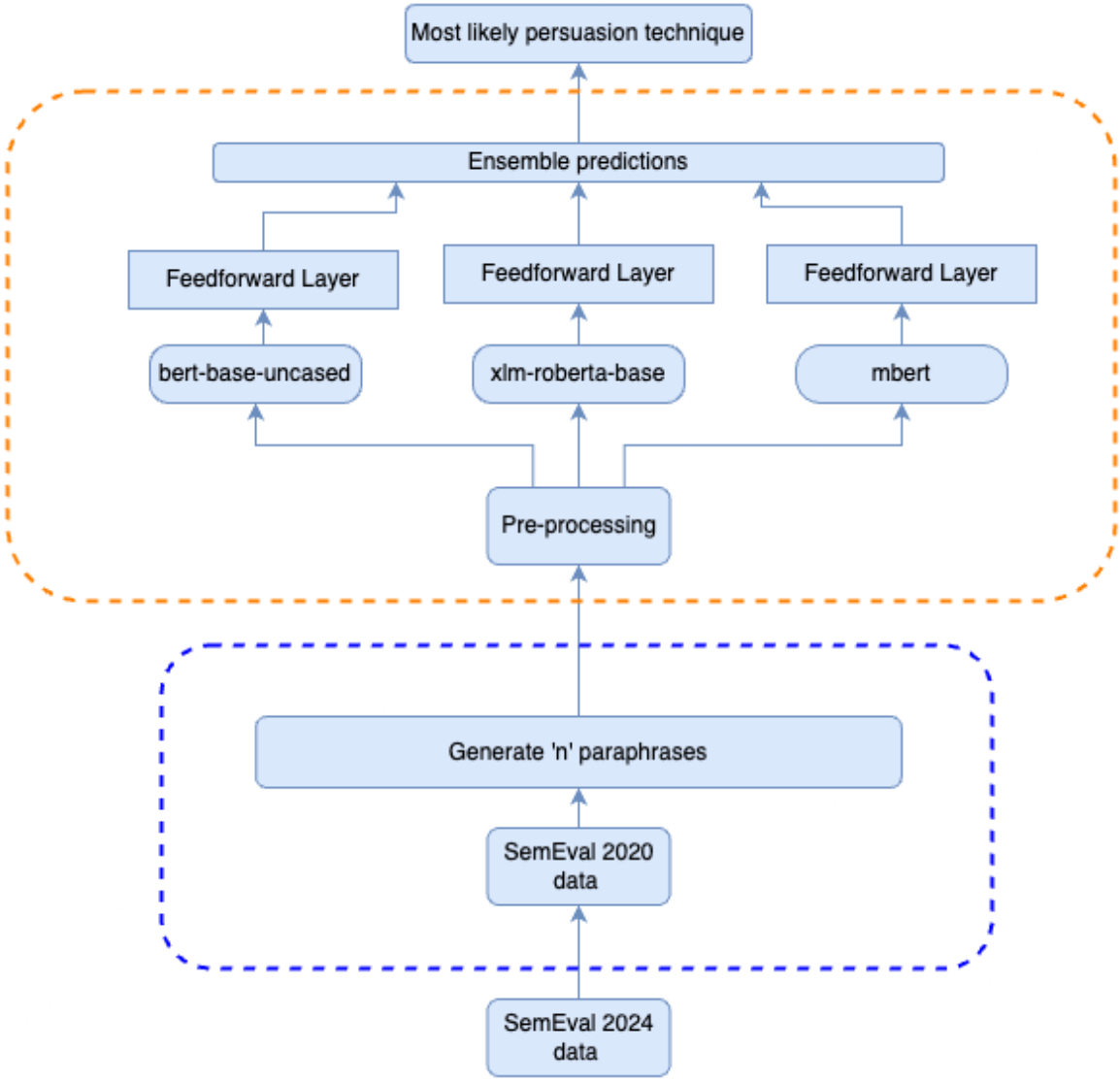

Figure 3.5: Schematic overview of our classification pipeline for the detection of persuasion techniques in the text of memes.

Despite the hierarchical organization of the persuasion techniques, we opted to predicting solely the technique names (leaf nodes in Figure 3.2) and not their ancestor nodes. However, to address the multi-label classification, we implemented thresholding in order to determine which techniques have a high enough score to be part of the output label set. We experimented with custom values for each technique with values ranging from 0.01 to 0.7 and picked the optimal values for each class based on the validation set. These thresholds were applied to the scores obtained after passing the logits of each class through a sigmoid function.

To handle the three surprise languages, during the official testing phase system, the model, trained only on English, would automatically translate the surprise language to English for our model's zero-shot predictions. This was inspired by the approach of (Costa, Hamilton, & Kosseim, 2023).

3.4.1 Data Augmentation

As Figure 3.3 shows, some persuasion techniques have very few samples (eg: *Red Herring* and *Straw Man* only have 59 and 62 instances respectively) in the SemEval 2024 training set. To mitigate the lack of data we took advantage of two data augmentation strategies:

(1) Adding related data from the Technique Classification subtask from SemEval 2020 task 11 (Da San Martino et al., 2020).

(2) Automatically generated paraphrases via a Large Language Model.

SemEval 2020 Data (`Comb-14k` dataset)

The Technique Classification (TC) subtask from the SemEval 2020 Task 11 (Da San Martino et al., 2020) provided a dataset with ≈7k instances annotated with the same guidelines as this year's. In contrast to the 2020 task, this year's challenge featured a revised set of techniques compared to the 2020 inventory. In the 2020 TC dataset, a few techniques were merged into a single category due to lack of data, resulting in a list of 14 techniques. In the current year, an expanded inventory of 20 techniques was employed. To ensure consistency between the two sets, we preprocessed the 2020 TC dataset by splitting techniques that had previously been merged. For example, we singled out *Bandwagon* and *Reductio ad Hitlerum*, which had been merged into a single technique in the SemEval 2020 TC dataset.

We considered two approaches to leverage the modified 2020 TC dataset. The initial option involved pre-training models on this dataset, followed by fine-tuning on the 2024 training data—an approach implemented by (Tian, Gui, Li, Yan, & Xiao, 2021). Another approach entailed combining both datasets and fine-tuning the models on this combined dataset. We chose the latter method because the two datasets covered different genres and a joint training approach would likely enable

the model to better adapt and grasp nuanced linguistic patterns across both.

We combined both datasets and fine-tuned models on this combined dataset. For easy reference in the rest of the thesis, we call the combined dataset `Comb-14k`. Figure 3.6 (orange + blue) shows the resulting distribution of the persuasion techniques in this combined dataset.

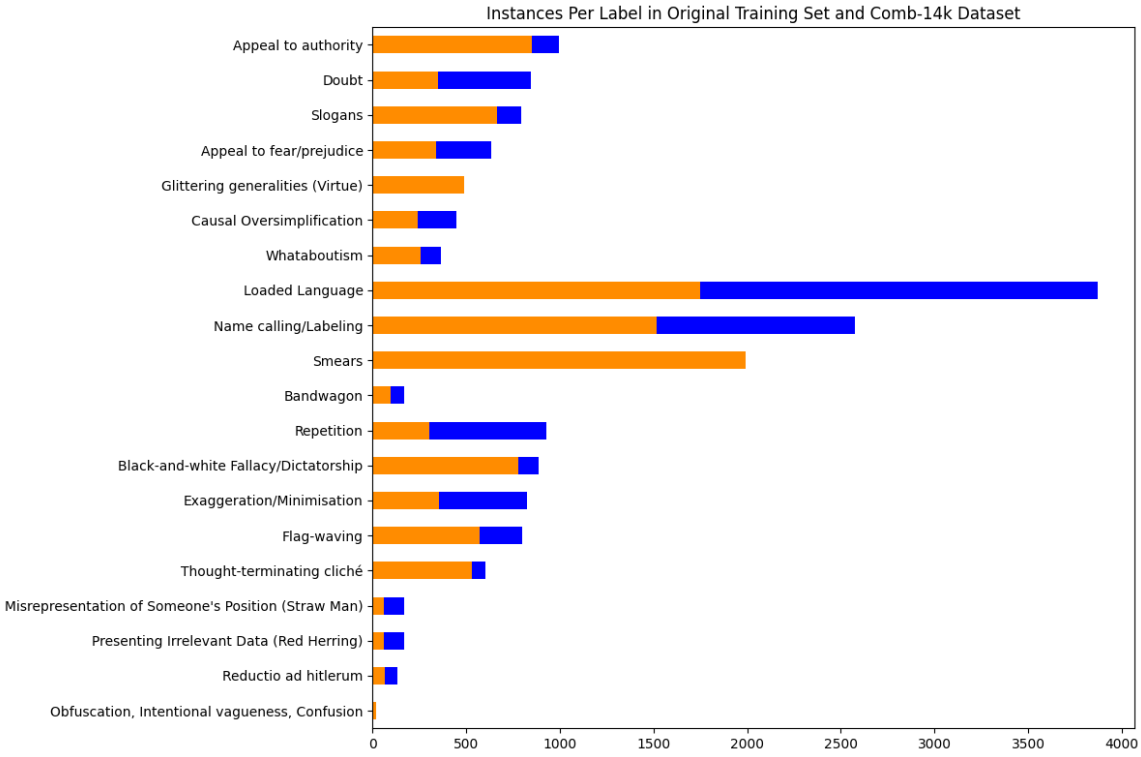

Figure 3.6: Distribution of the data for each persuasion technique in the original training set (in orange) and the `Comb-14k` dataset (in orange + blue)

Paraphrasing

Despite having almost doubled each class with the use of the 2020 TC dataset, some classes were still severely underrepresented; see Figure 3.5 (orange + blue). To address this, we augmented the dataset further by generating paraphrases for each instance. To generate paraphrases, we leveraged ChatGPT-3.5 turbo[1], setting the temperature to 0.7. This value aimed to introduce diversity in the paraphrases while maintaining relevance to the original instances. This allowed us to generate different datasets presented below.

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

`Para-n1` and `Para-n3`: For each instance in `Comb-14k`, we generated n paraphrases. We experimented with n=1 and n=3 leading to datasets of 28k and 52k respectively, which we call `Para-n1` and `Para-n3` respectively.

When validating our model (see Section 3.4.2) on the validation set given (500 instances), the overall hierarchical F-score showed an increase when training with these datasets and $n = 3$ seemed to perform better than $n = 1$.

However, a per-class analysis showed that not all classes benefited from the increase in support. For example, the persuasion technique *Bandwagon* increased its F1 from 0.17 to 0.29; whereas *Repetition* decreased its F1 from 0.56 to 0.31. We therefore identified the classes with improvement in F-score greater than 0.03 when using the `Para-n3` dataset compared to the `Comb-14k` dataset. These 8 techniques along with their increase in F-scores are shown in Figure 3.7. This set of techniques formed the basis for our subsequent strategy.

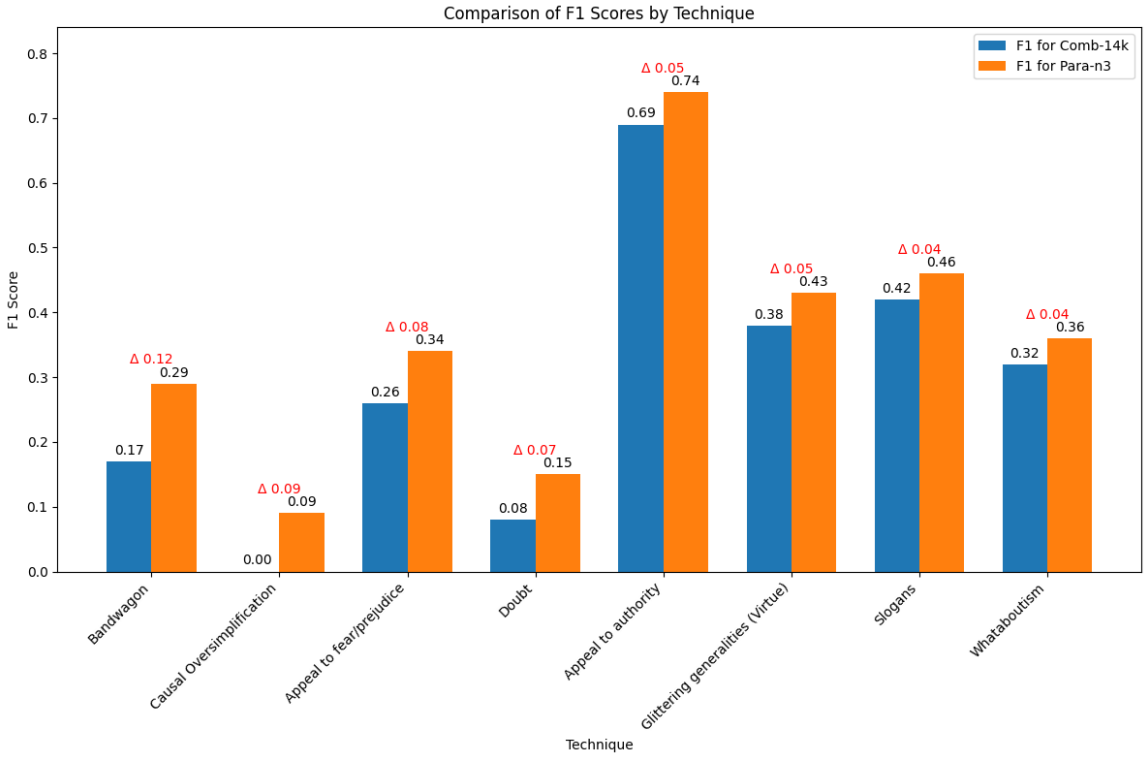

Figure 3.7: Techniques that showed an improvement in hierarchical F1 score with the validation set when using n=3 paraphrases (i.e. `Para-n3`) compared to `Comb-14k`.

`Para-Benef`: Since only 8 techniques seemed to benefit from the use of paraphrases, we created

a new augmented dataset by increasing the number of paraphrases only for these techniques. Specifically, let $\mathbf{B}$ be the set of 8 techniques that benefited from paraphrases (see Figure 3.7), for all data instances d in `Comb-14k` labeled with techniques $\mathbf{T} = \{t_1, t_2, \ldots t_n\}$ (where $n \leq 20$), for each $t_i \in \mathbf{B}$, we generated 10 paraphrases of d and labeled them with all techniques from $\mathbf{T} \cap \mathbf{B}$. This newly created dataset called `Para-Benef`, contained 54k instances.

Figure 3.8 shows the distribution of instances for each technique in the `Para-Benef` dataset (orange + blue + green), in comparison with the original training set and the `Comb-14k` dataset. As Figure 3.8 shows, the techniques such as *Bandwagon* and *Whataboutism* increased their representation significantly using this technique. Moreover, the datasets are severely imbalanced. Our next dataset therefore tried to address this issue.

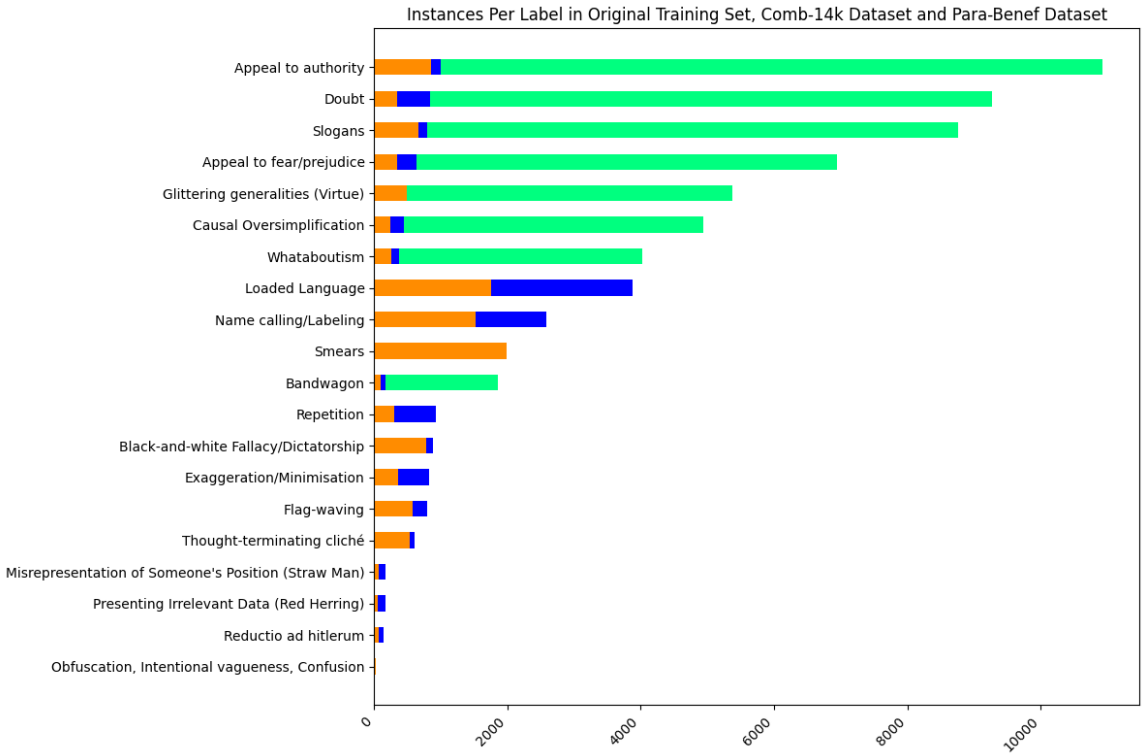

Figure 3.8: Distribution of the data for each persuasion technique in the original training set (in orange), the `Comb-14k` dataset (in orange + blue) and the `Para-Benef` dataset (in orange + blue + green)

`Para-Bal`: Our last dataset used our paraphrase generation strategy to address the dataset imbalance. We rectified the underrepresented classes in the initial training dataset by augmenting them with paraphrases. The most frequent three techniques—*Smears*, *Name-calling/Labelling*, and *Loaded Language* had 1990, 1750, 1518 samples respectively. We thus aimed at reaching similar number of instances for the other techniques. We balanced the dataset by generating batches of 5 paraphrases for each other technique to reach around 1500 instances. This newly created dataset called `Para-Bal` contained 49k instances (see Figure 3.9).

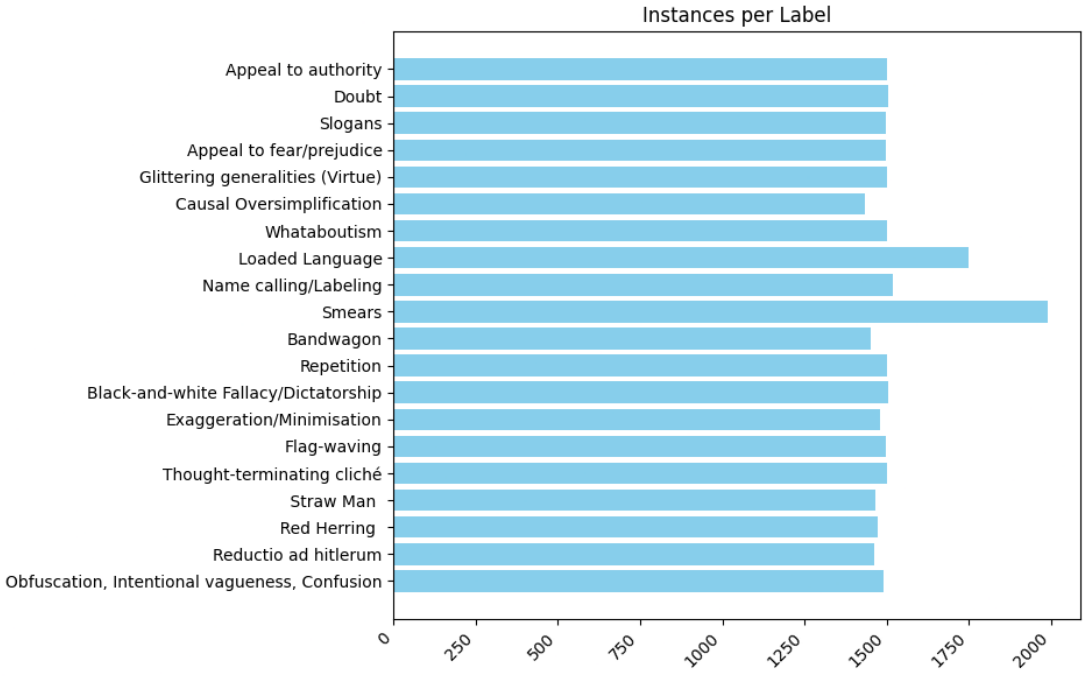

Figure 3.9: Distribution of techniques in the `Para-Bal` dataset.

3.4.2 Multi-label Classification

After creating the augmented datasets, we preprocessed them using standard tokenization, then proceeded to fine-tune the three distinct models in addition to an ensemble model, generated by averaging the predictions from all three models.

Additionally, we implemented thresholding in order to determine which techniques have a high enough score to be part of the output label set. We experimented with custom values for each of

Table 3.1: Hierarchical F1 scores of our models, when trained on different English-language datasets for both the validation and development sets.

Training Set	Models	Validation Set	Development Set
Original (7k)	BERT	0.42	0.43
	XLM-RoBERTa	0.48	0.49
	mBERT	0.48	0.48
	Ensemble Model	0.45	0.46
Comb-14k (14k)	BERT	0.52	0.55
	XLM-RoBERTa	0.53	0.54
	mBERT	0.53	0.54
	Ensemble Model	0.53	0.56
Para-n1 (28k)	BERT	0.55	0.57
	XLM-RoBERTa	0.57	0.54
	mBERT	0.50	0.53
	Ensemble Model	0.55	0.56
Para-n3 (52k)	BERT	0.54	0.55
	XLM-RoBERTa	0.54	0.54
	mBERT	0.54	0.55
	Ensemble Model	0.56	0.57
Para-Benef (54k)	BERT	0.48	0.51
	XLM-RoBERTa	0.54	0.55
	mBERT	0.51	0.53
	Ensemble Model	0.54	0.55
Para-Bal (49k)	BERT	0.54	0.58
	XLM-RoBERTa	0.58	0.59
	mBERT	0.53	0.55
	Ensemble Model	**0.59**	**0.61**

the techniques in order to address the data imbalance issue. We experimented with values ranging from 0.01 to 0.7 and picked the optimal values for each class based on the validation set (500 samples). These thresholds were applied to the scores obtained after passing the logits of each class through a sigmoid function. Table 3.1 shows the results of the validation with the optimal threshold for each class using the official SemEval scorer (Dimitrov et al., 2024), which uses hierarchical metrics. As Table 3.1 shows, the best model with the validation set was the ensemble trained on the Para-Bal dataset which reached an hierarchical F1 of 0.59. The ensemble model when trained on the Para-Benef dataset, performed worse (hierarchical F1 of 0.54 with the validation set) than the ones that used lesser number of paraphrases (Para-n1 and Para-n3). Since, the Para-Bal dataset was created after the SemEval competition, the ensemble, leveraging the collective insights of the three models, trained on the Para-n3 emerged as the most effective in enhancing the overall

system performance during the competition. Based on our results in the official leaderboard with the development set and validation results shown in Table 3.1, we chose to submit the ensemble model trained on the `Para-n3` dataset as it gave the best results with both the validation and the development set. Surprisingly, `Para-Benef` which contained 10 paraphrases for the benefited techniques did not perform better than using only 3 paraphrases for all techniques (`Para-n3`). This suggests that the excessive inclusion of paraphrases from a different distribution (memes versus news) may have led to too much noise in the data.

To deal with the surprise languages, the system was set up to automatically translate the datasets to English for our model's zero-shot predictions. This was inspired by the approach of Costa et al. (2023). The primary reason for employing automatic translation is the model's limited training on non-English data. Given that our model was fine-tuned exclusively on English datasets, directly applying it to non-English text would likely result in poor performance due to the lack of exposure to these languages. By translating the test datasets into English, we leveraged the model's strong performance in English to make predictions on content originally in other languages. The English test data was used as given.

3.5 Experimental Setup

The system pipeline code was implemented in PyTorch. The pre-trained models BERT, XLM-RoBERTa, and mBERT and their tokenizers[2] were sourced from Hugging Face.

All models were trained for 10 epochs using the Adam optimizer with a learning rate of 2e-5. Batch sizes varied with BERT utilizing 128, and XLM-RoBERTa and mBERT using 64. A final feedforward layer with 20 logits (equal to the number of persuasion techniques) was added to each model. The Binary Cross Entropy with logits served as the loss function, with one-hot encoding applied to the true labels. For prediction, a sigmoid activation function was used on the logits, followed by thresholding. The ensemble model used an unweighted average of all predictions from the three individual models. The ChatGPT-3.5 turbo[3] API with a temperature set to 0.7 was used for paraphrase generation.

[2] https://huggingface.co/docs/transformers/en/main_classes/tokenizer
[3] https://platform.openai.com/docs/models/gpt-3-5-turbo

During the testing phase, the datasets in the surprise languages were automatically translated to English for our model's zero-shot predictions using the deep-translator API[4].

3.6 Results and Analysis

3.6.1 Official SemEval 2024 Results

For our official submission to SemEval 2024, the `Para-Bal` dataset had not been created yet; hence our official results are based on the ensemble model trained on the union of `Para-n3` and the development set (1k samples), for a total of 53k samples. The three surprise languages turned out to be Bulgarian, North Macedonian and Arabic. The test set contained 1500 samples for English, 426 samples for Bulgarian, 259 samples for North Macedonian and 100 samples for Arabic. The official results of our system are shown in Figure 3.10 and Table 3.2, along with a baseline score that assigns the most frequent persuasion technique (i.e. *Smears*) to all instances, and the score obtained by the best performing systems for each language (D. Li et al., 2024; Wunderle et al., 2024). As Table 3.2 shows, although our ensemble model did not reach the top performance for English (0.57 versus 0.75), it performed better than the baseline in all languages except Arabic, where the improvement was not significant.

[4] https://pypi.org/project/deep-translator/

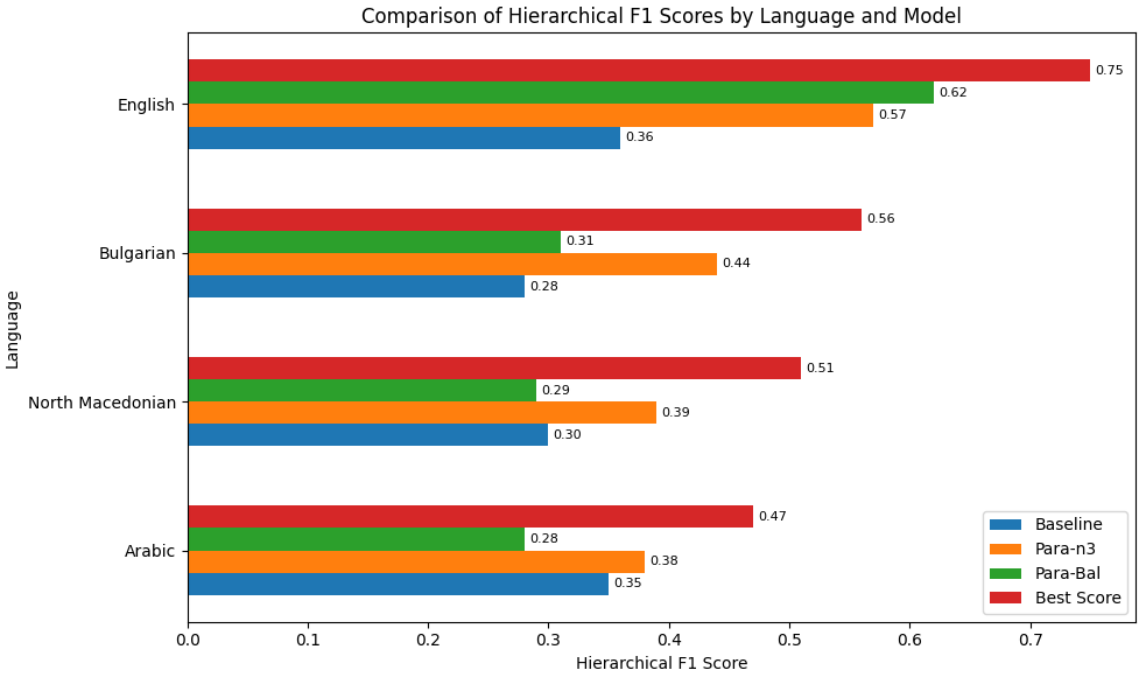

Figure 3.10: Comparison of the final hierarchical F1 scores obtained by our official SemEval 2024 model trained with `Para-n3`, the model trained with `Para-Bal`, the best corresponding models (D. Li et al. (2024) for English, Wunderle et al. (2024) for Bulgarian, North Macedonian and Arabic) in the shared task and the baseline in each given language.

Overall, we stood at 22nd out of 33 participants for English, 12th out of 20 for Bulgarian, 11th out of 20 for North Macedonian and 11th out of 17 for Arabic.

Language	**Baseline**	**Our Score**	**Best Score**
English	0.36865	0.57827	0.75427
Bulgarian	0.28377	0.44917	0.56833
North Macedonian	0.30692	0.39471	0.51244
Arabic	0.35897	0.38070	0.47593

Table 3.2: Comparison of the final hierarchical F1 scores obtained by our official SemEval 2024 model, the best corresponding classification system in the shared task and the baseline in each given language.

3.6.2 Post Shared Task Results

Using the same testing protocol, we reproduced the results of SemEval 2024 using the model trained with the balanced training dataset (`Para-Bal`). The results displayed in Figure 3.10 indicate an improvement in score with the English test set (0.62 versus 0.57). This again confirms

the importance of a balanced dataset, and paraphrases based on the same distribution as the original texts. Indeed, although `Para-n3` is larger than `Para-Bal` (52k versus 49k), it is not balanced and contains paraphrases of instances from different genres (memes and news). However, surprisingly, the performance enhancement when using `Para-Bal` is not observed for the zero-shot classification of the surprise languages whose performance dropped significantly. For these languages, a larger training set, even with noisy out-of-distribution instances, leads to better results possibly due to the noise introduced by the automatic translation itself.

Compared to the other approaches at the 2024 edition of SemEval Task 4, the top performing team overall, OtterlyObsessedWithSemantics (Wunderle et al., 2024) created a custom classification layer for a large language model (LLaMA 2) that used multiple layers to reflect a hierarchy. This setup allowed decisions made at broader levels to guide more detailed predictions, helping the model classify persuasion techniques more accurately. They optimized the model's performance by systematically exploring different hyperparameters through grid-search. In addition, similarly to our approach, during the testing phase, they translated all the surprise language datasets into English. The second best team (D. Li et al., 2024), developed a system using Chain-of-Thought based data augmentation methods (using GPT-3.5), in-domain pre-training and ensemble strategy that combined the strengths of both RoBERTa and DeBERTa models.

3.7 Chapter Summary

This chapter described our approach to identifying persuasion techniques within meme texts as part of Subtask 1 of SemEval 2024 Task 4 (Dimitrov et al., 2024). The task required multi-label classification of 20 persuasion techniques in meme texts across multiple languages, including three surprise languages for zero-shot testing. We presented the dataset used, highlighting its imbalanced nature, particularly the underrepresentation of certain techniques, which we addressed through paraphrase-based data augmentation strategies.

Our methodology involved fine-tuning three pre-trained language models—BERT, XLM-RoBERTa, and mBERT—on the training dataset, with predictions averaged to form an ensemble model. To optimize the multi-label classification performance, we implemented custom thresholding for each

persuasion technique. Despite the hierarchical nature of the techniques, our focus was on predicting only the leaf nodes, balancing performance with model complexity.

Data augmentation played a crucial role in enhancing the dataset, using both existing datasets from prior tasks and paraphrased instances generated via the ChatGPT-3.5 turbo API. We iteratively refined our paraphrasing strategy, creating multiple datasets. Our findings indicated that while paraphrasing generally improved performance, excessive augmentation could introduce noise, emphasizing the need for careful control.

During the testing phase, we evaluated the model's zero-shot performance on multilingual meme texts, automatically translating non-English datasets to English. This approach allowed the model to leverage its English training while predicting on other languages. However, the translation process introduced some noise, affecting performance on non-English texts.

Our analysis during the task included comparing our system's performance with top-performing models in the shared task. Although our approach was not the top performer, it effectively managed the complexity of multi-label persuasion detection in meme texts, particularly in a zero-shot multilingual setting. Post-task, we focused on creating a new balanced dataset, which further enhanced our system's capability.

In the next chapter, we will discuss the identification of persuasion techniques within multimodal meme content, integrating both textual and visual data.

Chapter 4

Multimodal Multi-label Classification of Persuasion Techniques in Memes

4.1 Introduction

This chapter provides a description of our methodology for identifying persuasion techniques within meme texts and images. Our goal is to investigate how various multimodal techniques can be employed to detect persuasive strategies used in memes.

We focus on the use of state-of-the-art neural architectures that combine both textual and visual data to understand how these modalities work together to convey persuasive messages. Specifically, we discuss two approaches. The first is an early fusion approach (see Section 2.4) where the image embeddings produced by CLIP (see Section 2.2) and the text embeddings generated by XLM-Roberta (see Section 2.2) are concatenated and passed through a Multi-Layer Perceptron (MLP). This strategy allows us to jointly process both image and text features at an early stage, enabling the model to learn interactions between the modalities and effectively capture multimodal persuasive cues. The second approach we experimented with is based on a cross-modal alignment approach (see Section 2.4) using VisualBERT (see Section 2.2), where the image and text encodings (from ResNet (see Section 2.2) and BERT (see Section 2.2), respectively) are handled separately and aligned through VisualBERT's cross-attention mechanism. This method was explored to assess whether a more fine-grained alignment between the modalities could improve performance over

the early fusion approach. We find that, surprisingly, this approach did not lead to better results. One possible reason could be overfitting due to the complexity of aligning image and text features separately through the cross-attention mechanism in VisualBERT. Additionally, the increased complexity may have led to difficulties in optimizing the model effectively, which resulted in its underperformance compared to the simpler early fusion approach. Our early fusion approach using CLIP and XLM-Roberta demonstrated strong performance, underscoring the importance of joint multimodal processing in effectively capturing persuasive techniques.

Throughout this chapter, we detail the steps involved in preprocessing the data, the model architecture, fusion strategies, and results used to assess our approach. We also provide an analysis of the performance of multimodal models compared to single-modality approaches (Text-Only and Image-Only), offering insights into the advantages of leveraging both modalities in identifying persuasion techniques in memes.

4.2 About the Task

As indicated in Section 3.2, the SemEval-2024 shared Task 4 (Dimitrov et al., 2024) aimed at understanding how memes employ various persuasion techniques to influence user perspectives. Subtask 1 focused solely on the analysis of textual content, involving 20 labels (see Chapter 3), while Subtask 2 integrated both text and visual elements and expanded the classification to 22 labels. Although the training dataset was provided in English, the evaluation for both subtasks required testing the model's zero-shot performance across three surprise test languages along with an additional English dataset. The testing phase was intended to assess the model's generalization capability to these languages without explicit training. This chapter describes our work on Subtask 2, which focuses on detecting 22 hierarchically organized persuasion techniques within the textual and visual content of memes.

4.3 Datasets

For Subtask 2, the goal was to categorize both the textual and visual content of memes into zero or several persuasion techniques. The inventory of techniques was expanded to include 22 labels,

with two additional techniques, *Transfer* and *Appeal to (Strong) Emotions*, added to the original set. These 22 techniques were also structured in a directed acyclic graph, rendering the task a hierarchical multi-label classification problem.

For example, given the training instance shown in Figure 4.1 along with the image in Figure 4.2, the model needs to learn that the text *HAPPY NEW YEAR FROM PRESIDENT DONALD J. TRUMP* should be labelled with three techniques provided in the labels field.


```
{
    "id": "70817",
    "text": "HAPPY NEW YEAR FROM PRESIDENT DONALD J. TRUMP",
    "image": "prop_meme_6647.png",
    "labels": [
       "Transfer",
       "Flag-waving",
       "Slogans"
    ],
    "link": "null"
 }
```


Figure 4.1: A sample training instance which has both the text and the image modalities. The instance is labelled with three techniques, *Transfer*, *Flag-waving*, *Slogans*. The image `prop_meme_6647.png` is provided in Figure 4.2.

The training (7k samples), validation (500 samples) and development (1k samples) sets included only English texts; whereas the test set was multilingual with 1500 samples for English, 426 samples for Bulgarian, 259 samples for North Macedonian and 120 samples for Arabic. All datasets were the provided in the form of JSON files and the images were in `png` format.

Figure 4.2: Image corresponding to the sample instance in Figure 4.1. (Image source: `prop_meme_6647.png` from the training dataset)

4.4 Preprocessing

This section provides a description of our methodology for detecting persuasion techniques in memes by fine-tuning multimodal models that leverage both image and text data. Our approach combines early fusion and cross-modal alignment strategies, applying state-of-the-art models such as CLIP, XLM-Roberta, and VisualBERT, fine-tuned to the task of persuasion detection.

Effective preprocessing is essential to ensure that both image and text data are in formats compatible with the models.

4.4.1 Image Preprocessing

Given the diverse nature of meme images, which often include a mixture of text, illustrations, and photographs, our preprocessing pipeline aims to standardize the visual input. For the image preprocessing, we applied transformations tailored to the input requirements of CLIP and ResNet using Pytorch's `torchvison.transforms`[1].

- CLIP: The images were resized to 336×336 pixels using bicubic interpolation and then center-cropped to the same dimensions. The images were converted to RGB format before being normalized into tensors. These transformations are consistent with the preprocessing used during CLIP's pretraining to ensure compatibility with the pretrained model.
- ResNet: The images were resized to 224×224 pixels and center-cropped to maintain this size. Following this, the images were converted to RGB format and normalized before converting them into tensors.

These preprocessing steps ensure that the input images are appropriately formatted for their respective models, maintaining consistency with the preprocessing pipelines used during the models' pretraining.

4.5 Proposed Model Architectures

We explore four distinct multimodal model architectures to detect persuasion techniques within memes.

(1) **Model-Early** follows an early fusion approach, combining embeddings from both text (XLM-Roberta) and image (CLIP) modalities, which are then processed through a MLP.

(2) **Model-Cross** adopts a cross-modal alignment strategy using VisualBERT, where image features (ResNet) and text features (BERT) are aligned via VisualBERT's cross-attention layers.

(3) **Model-Image-Only** focuses exclusively on the image modality, fine-tuning CLIP embeddings passed through an MLP to assess the visual content's role in persuasion.

[1] https://pytorch.org/vision/0.9/transforms.html

(4) **Model-Text-Only** handles only textual content, using XLM-Roberta fine-tuned for detecting persuasive strategies in meme texts.

For the remainder of this thesis, we will refer to these architectures as `M-Early`, `M-Cross`, `M-Image-Only`, and `M-Text-Only`. A summary of these models is presented in Table 4.1.

Figures 4.3 and 4.4 illustrate the architectures for `M-Early` and `M-Cross`, respectively. `M-Image-Only` follows the same architecture as `M-Early` but without the text modality, and `M-Text-Only` uses the same architecture as `M-Early` but without the image modality.

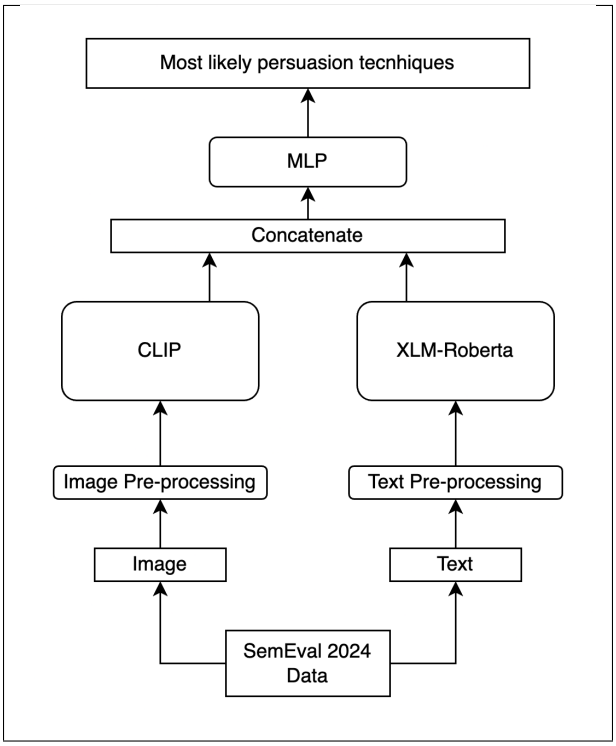

Figure 4.3: Architecture of `M-Early`: Early fusion of image and text embeddings (CLIP + XLM-Roberta) processed through an MLP.

Model	Approach	Text Modality	Image Modality
`M-Early`	Early Fusion + MLP	XLM-Roberta	CLIP
`M-Cross`	Cross-Modal Alignment (VisualBERT) + MLP	BERT	ResNet
`M-Image-Only`	Image-Only + MLP	-	CLIP
`M-Text-Only`	Text-Only + MLP	XLM-Roberta	-

Table 4.1: Summary of the Multimodal Model Architectures used for Subtask 2

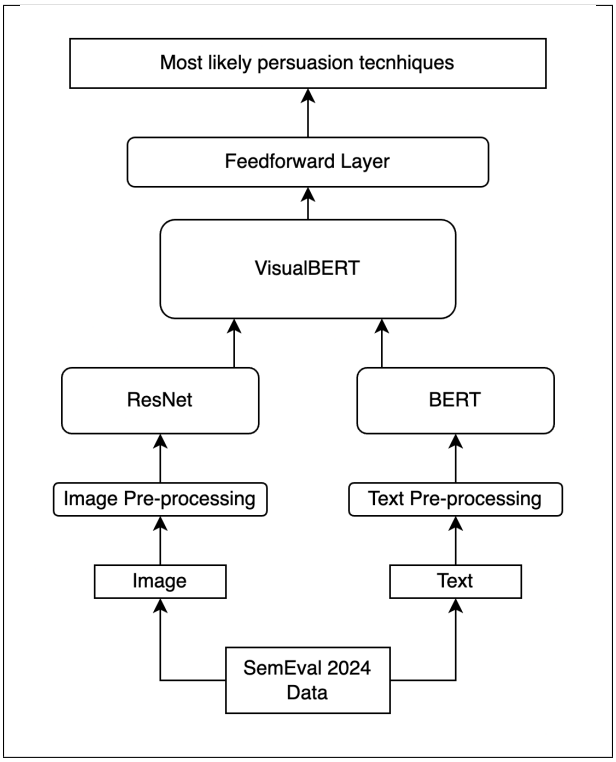

Figure 4.4: Architecture of `M-Cross`: Cross-modal alignment of image (ResNet) and text (BERT) embeddings using VisualBERT's cross-attention mechanism.

4.5.1 Model-Early : Early Fusion

The persuasive nature of memes often emerges from a combination of visual cues and textual rhetoric. For example, an image of a popular figure paired with ironic text may serve as a powerful persuasive device. By combining the visual and textual modalities early in the process, we can capture the interaction more effectively. The assumption here is that joint representations of image and text allow the model to learn the intricate interactions between the two, leading to more accurate persuasion detection.

To achieve this, we use an early fusion approach by concatenating the embeddings from CLIP and XLM-Roberta. CLIP, trained on a large corpus of image-text pairs (see Section 2.2, excels at aligning visual and textual concepts, making it suitable for memes where the visual content carries persuasive elements. XLM-Roberta complements this by offering robust textual embeddings, especially in multilingual contexts. These concatenated embeddings are then passed through a Multi-Layer Perceptron (MLP) to process the joint representations, enabling the model to learn from both

modalities in tandem. Fine-tuning allows the models to adapt to the specific task of persuasion detection and better capture the nuances of meme content.

4.5.2 Model-Cross : Cross-Modal Alignment

Persuasion often relies on the interplay between text and image in memes. A seemingly neutral image may take on a persuasive tone when paired with a specific caption, or vice versa. Cross-modal alignment allows the model to detect instances where the image and text reinforce, contradict, or complement each other, which is a common technique in memes.

For this approach, we extract visual features using ResNet and textual features using BERT. These are then passed into VisualBERT, which uses cross-attention layers to align the two modalities. VisualBERT allows the model to learn intricate interactions between the image regions and corresponding textual phrases. By fine-tuning VisualBERT, we ensure that the model captures the detailed relationships essential for detecting persuasion techniques in memes.

For comparision points we also evaluated single-modality models which are described in Sections 4.5.3 and 4.5.4.

4.5.3 Model-Image-Only : Image-Only Approach

Visual content in memes often evokes emotions, sets a tone, or invokes cultural references. Even without text, images alone can carry persuasive intent, such as an image invoking nostalgia or anger. Detecting these subtle visual cues is essential for identifying persuasive elements within memes.

To isolate the image modality, we use CLIP, which is particularly well-suited to understanding visual concepts that may carry implicit messages or emotions. The image embeddings generated by CLIP are passed through an MLP to process only the visual content, allowing the model to focus on how much persuasive content is conveyed through images alone.

4.5.4 Model-Text-Only : Text-Only Approach

The text in memes often serves as a direct or indirect argumentative tool, leveraging rhetorical strategies like bias, humor, sarcasm, or propaganda to persuade the viewer. Even without accompanying visuals, the textual content can carry significant persuasive power.

In this model, we focus exclusively on the textual content by fine-tuning XLM-Roberta to detect persuasion techniques. The model captures the subtleties of language, such as rhetorical devices or humor, that are crucial for persuasion. The text embeddings are processed through an MLP, allowing the model to focus entirely on the linguistic aspects of memes to detect persuasive elements effectively.

4.6 Zero-Shot Experiments with GPT-4

In addition to the fine-tuned models (Model-Early to Model-Text-Only), we conducted zero-shot experiments using GPT-4 to evaluate how well it can detect persuasion techniques in memes without any prior training on the dataset. These experiments were designed to assess GPT-4's performance across four different prompt settings, each increasing in complexity, in a manner similar to Chain-of-Thought prompting (Wei et al. (2023)), where the model is guided to reason step by step through progressively more detailed prompts.

For cost reasons, the experiments were performed only on a subset of the dataset extracted from the development set, consisting of three mutually exclusive random subsets of 100 samples each. We used these subsets to measure the average and the standard deviation of GPT-4's performance across different prompt settings, ensuring a robust assessment of the model's stability and accuracy.

4.6.1 Prompt Settings

We used four progressively more detailed prompt settings, each designed to test the limits of GPT-4's reasoning abilities in a zero-shot context. The prompt settings are intended to simulate varying levels of reasoning complexity, gradually guiding the model toward more accurate predictions. These prompts range from simply providing a list of persuasion techniques to including definitions and examples for each technique. Detailed descriptions of these prompts can be found in Appendix B, which includes both the text of the prompts and the API query format used in our experiments.

Prompt 1: List of Labels

In this simplest setting, GPT-4 was provided with only a list of the possible persuasion techniques (see Appendix B.2). The task for the model was to infer which label or labels best fit the meme based solely on the list.

This prompt serves as a basic zero-shot classification task where GPT-4 must rely on its prior knowledge and understanding of the persuasion techniques, without any further guidance. The model is expected to perform basic pattern matching between the meme and its internal representations of the techniques, but it lacks context, making it prone to confusion or ambiguity in more complex cases.

Prompt 2: List of Labels + Definitions

In this setting, the list of persuasion techniques was accompanied by their definitions (see Appendix B.3). This added layer of explanation clarifies the meaning and scope of each technique, helping GPT-4 to better disambiguate between similar techniques.

By providing definitions, we enable GPT-4 to understand the nuances of each persuasion technique. This setting reduces the likelihood of misclassification that may arise from ambiguous or closely related labels. The model can draw on the explicit descriptions of the techniques to refine its predictions, leading to more accurate results compared to the label-Only setting.

Prompt 3: List of Labels + Examples

In the third setting, the model was given a list of the persuasion techniques along with examples of each (see Appendix B.4). These examples were designed to showcase clear, unambiguous instances of each technique being used in context.

The rationale for providing examples was to help GPT-4 contextualize each technique, demonstrating how they are used in real-world scenarios. This setup encourages a form of implicit Chain-of-Thought reasoning, where the model can compare the meme against known examples and reason by analogy. The use of concrete examples is expected to help the model recognize more subtle or abstract instances of persuasion techniques.

Prompt 4: List of Labels, Definitions, and Examples

The most complex prompt included a list of persuasion techniques, definitions, and specific examples for each technique (see Appendix B.5). This comprehensive prompt provided GPT-4 with a full conceptual and practical understanding of the task, combining the explanatory power of the definitions with the clarity of the examples.

This setting aligns most closely with the principles of Chain-of-Thought prompting. By giving the model detailed information and concrete examples, we enable it to engage in a deeper reasoning process. The combination of definitions and examples allows GPT-4 to link the theoretical understanding of each technique with practical applications, thereby improving its ability to generalize to unseen memes. This setting is expected to yield the best performance as it equips GPT-4 with all the tools necessary to accurately detect persuasion techniques.

While the prompt settings above do not explicitly instruct GPT-4 to perform Chain-of-Thought reasoning, the increasing complexity of the prompts encourages the model to engage in a more structured reasoning process. In Chain-of-Thought prompting, the model is guided through intermediate reasoning steps, leading to more accurate and coherent outputs. Here, the addition of definitions and examples serves a similar function by providing the model with intermediate "checkpoints" that it can use to refine its understanding of the task.

For example, in the most complex prompt (see Prompt 4), GPT-4 can first reason about the definition of a technique and then compare the meme to the provided examples. This layered reasoning process mimics the multi-step thought process that a human might engage in when identifying persuasion techniques, thus improving the model's performance in detecting subtle and context-dependent instances of persuasion.

For the remainder of this thesis, we will refer to GPT-4's zero-shot performance as `M-GPT4`. Each of the four prompt settings will be referred to as `M-GPT4-p1`, `M-GPT4-p2`, `M-GPT4-p3`, and `M-GPT4-p4`, respectively, corresponding to the specific prompts used in those settings.

4.7 Experimental Setup

For all fine-tuned models (`M-Early` to `M-Text-Only`), fine-tuning allows to adapt pretrained models (CLIP, XLM-Roberta, BERT, and VisualBERT) to the specific task of detecting persuasion techniques. By adjusting their weights based on the meme dataset, we ensure that the models become more sensitive to the patterns that are characteristic of persuasive memes, rather than generic visual or textual features.

Our experiments with the validation set showed that using the Adam optimizer[2] without weight decay provided better performance compared to using weight decay. We hypothesize that weight decay may have interfered with the learning process in this case, potentially penalizing the model for relying on important features that are more domain-specific in the context of memes.

We fine-tuned all models on the Subtask 2 dataset. The models were trained for 20 epochs with a batch size of 32. A final feedforward layer with 22 logits (equal to the number of persuasion techniques) was added to each model. The Binary Cross Entropy with logits served as the loss function, with one-hot encoding applied to the true labels. For prediction, a sigmoid activation function was used on the logits, followed by thresholding.

Following the task organizers' specifications, we use hierarchical precision, recall, and F1-score to evaluate model performance. We first tested the models on the development set before selecting the best-performing one for the test set evaluation. During the testing phase, the text in the datasets from the surprise languages was automatically translated to English for our model's zero-shot predictions, using the deep-translator API[3]. However, the images were used in their original form without translation. This approach was feasible because the CLIP model, which we used to extract image embeddings, is trained on a diverse range of image-text pairs and is robust to the language of the text inscribed in images. Additionally, even when ResNet is used for extracting image embeddings, the textual content in the images does not need translation, as the model focuses on visual features rather than linguistic content. This allows the models to capture the persuasive elements conveyed through visual cues regardless of language differences.

[2] https://pytorch.org/docs/stable/generated/torch.optim.Adam.html
[3] https://pypi.org/project/deep-translator/

The zero-shot experiments (M-GPT4) were conducted on three mutually exclusive random subsets of 100 memes each, drawn from the development set. These subsets were used to evaluate GPT-4's performance under different prompt settings. The use of mutually exclusive subsets ensures that the model's performance is evaluated across a diverse range of meme types, allowing us to assess the generalizability of its predictions.

To enable GPT-4 to process images as part of its multimodal capabilities, we use the GPT-4 API[4], which requires that images be encoded in base64 format. This encoding transforms the images into a format that GPT-4 can interpret alongside the text, enabling the model to reason about both visual and textual content in a zero-shot manner. Appendix B.1 shows the query to the API.

We measured performance using hierarchical precision, recall, and F1-score metrics, consistent with the evaluation metrics used in the fine-tuning experiments. Additionally, we calculate the average and standard deviation of the model's predictions across the three subsets to assess the stability and consistency of GPT-4's performance. By analyzing these statistics, we assess whether increasing the complexity of the prompt leads to more reliable predictions.

To contextualize GPT-4's performance, we also compared it against the results of our best fine-tuned model (M-Early) on the same three mutually exclusive datasets. This comparison provides insight into the strengths and limitations of zero-shot learning compared to task-specific fine-tuning. The fine-tuned model, having been trained specifically on the task of persuasion detection, serves as a baseline, offering a direct comparison with GPT-4, which has broad, generalized knowledge but no specific training on this task.

4.8 Results and Analysis of the Fine-Tuned Models (Models 1 to 4

4.8.1 Results

The results in Table 4.2 show the performance of the fine-tuned models. As the table shows, there are clear differences in performance across the various architectures. M-Early (XLM-Roberta + CLIP + MLP), which follows an early fusion strategy, achieved the highest hierarchical

[4] https://platform.openai.com/docs/models/gpt-4o

F1 score, with values of 0.63 on the validation set and 0.65 on the development set. This strong performance can be attributed to the combination of powerful image and text encoders and the ability of the MLP to learn joint representations that capture both modalities effectively.

Models	**Validation Set**	**Development Set**
`M-Early` : XLM-Roberta + CLIP + MLP	0.63	**0.65**
`M-Cross` : BERT + ResNet + VisualBERT	0.39	0.40
`M-Image-Only`: CLIP + MLP	0.62	0.64
`M-Text-Only` : XLM-Roberta + MLP	0.50	0.54

Table 4.2: Hierarchical F1 scores of our models, for both the validation and development set

`M-Image-Only` (CLIP + MLP), which uses only image features from CLIP, performed slightly lower than the XLM-Roberta + CLIP model, but still achieved strong results, with hierarchical F1 scores of 0.62 and 0.64 on the validation and development sets, respectively. This result highlights the strength of CLIP's image embeddings, suggesting that much of the persuasive content in memes is conveyed through visual elements alone.

`M-Text-Only` (XLM-Roberta + MLP), which only processes textual content, performed moderately well, with hierarchical F1 scores of 0.5 on the validation set and 0.54 on the development set. The lower scores compared to the multimodal approaches indicate that textual content, while important, does not fully capture the persuasive techniques in memes on its own. This supports the idea that persuasive messages in memes are often a combination of text and imagery, and a text-Only approach may miss out on the context provided by the visual modality.

Surprisingly, `M-Cross` (BERT + ResNet + VisualBERT), which employs cross-modal alignment through VisualBERT's attention mechanism, performed the lowest across all models, with hierarchical F1 scores of 0.39 on the validation set and 0.4 on the development set. This result suggests that ResNet and BERT embeddings may not be well-aligned for the task of persuasion detection, leading to suboptimal cross-modal alignment. VisualBERT's attention mechanism, while effective for certain more factual tasks like visual question answering, may struggle to capture the more abstract and nuanced interactions between images and text that are characteristic of persuasive techniques in memes. The complexity of aligning the two modalities might have also led to overfitting, further contributing to the poor results.

4.8.2 Analysis

The superior performance of `M-Early` confirms our hypothesis that early fusion, where both visual and textual features are combined before classification, allows for better learning of multimodal relationships in the context of persuasion detection. The strength of this model lies in its ability to leverage the complementary nature of image and text. The image encodings from CLIP effectively capture visual features that convey emotions, symbolism, or humor, while the textual encodings from XLM-Roberta provide insights into the rhetorical structure and language used in the meme text. This joint representation leads to a more complete understanding of the persuasive message.

`M-Image-Only`'s high performance demonstrates that CLIP's image embeddings are robust enough to handle the task on their own in many cases. Memes often use powerful imagery to evoke strong emotional responses or convey implicit messages, and CLIP's pretrained knowledge on image-text pairs enables it to understand the context of these visual elements effectively, even without textual input.

`M-Text-Only`'s lower performance indicates that text alone is not sufficient for optimal performance in persuasion detection within memes. While XLM-Roberta is adept at understanding complex language, memes often rely on visual context to convey the full persuasive intent. Thus, the lack of visual information limits the model's ability to fully interpret the persuasive elements at play.

The poor performance of `M-Cross` was surprising, however it may indicate that simply using cross-attention to align image and text features from ResNet and BERT may not be sufficient for this task. The likely reason for this underperformance is the misalignment between the embeddings produced by ResNet and BERT, which may not be fully compatible for fine-grained alignment through VisualBERT's cross-attention mechanism. Additionally, ResNet's focus on lower-level visual features may hinder its ability to capture the abstract, high-level visual elements necessary for detecting persuasion techniques.

4.9 Results and Analysis of GPT-4 Prompting

4.9.1 Results

Table 4.3 presents the hierarchical F1 scores of our best fine-tuned model, `M-Early` alongside GPT-4's zero-shot performance (`M-GPT4`) across four different prompt settings. The results are reported for three mutually exclusive random subsets of 100 memes drawn from the development set.

GPT-4 Performance Across Prompts:

- **Prompt 1** (Labels Only): GPT-4 achieved hierarchical F1 scores of 0.49, 0.53, and 0.49 across the three subsets. This basic prompt, which only provided a list of labels, led to moderate performance, as GPT-4 had to rely solely on its general understanding of the labels without further guidance.
- **Prompt 2** (Labels + Definitions): When definitions were provided alongside the labels, GPT-4's performance improved slightly, with F1 scores rising to 0.54, 0.56, and 0.54 across the subsets. The definitions helped clarify the meaning of each persuasion technique, allowing GPT-4 to better distinguish between them.
- **Prompt 3** (Labels + Examples): With examples provided instead of definitions, GPT-4's F1 scores were 0.53, 0.55, and 0.52. Although examples helped illustrate the application of the techniques, the model's performance did not significantly improve over the previous prompt. This suggests that while examples aid contextual understanding, they might not offer the same conceptual clarity as definitions.
- **Prompt 4** (Labels + Definitions + Examples): This most detailed prompt configuration, which combined both definitions and examples, resulted in GPT-4's highest scores across the three subsets, with F1 scores of 0.56, 0.57, and 0.54. The richer context provided by this prompt helped GPT-4 reason more effectively about the persuasive techniques present in the memes.

Performance of the Fine-Tuned Model: The fine-tuned XLM-Roberta + CLIP + MLP model consistently outperformed GPT-4 across all three subsets, achieving hierarchical F1 scores of

Prompts	Models	Random Set 1	Random Set 2	Random Set 3
Prompt 1	M-GPT4-p1	0.49	0.53	0.49
Prompt 2	M-GPT4-p2	0.54	0.56	0.54
Prompt 3	M-GPT4-p3	0.53	0.55	0.52
Prompt 4	M-GPT4-p4	**0.56**	**0.57**	**0.54**
No Prompt	M-Early	**0.66**	**0.69**	**0.68**

Table 4.3: Comparison of Hierarchical F1 Scores Between GPT-4 (Zero-Shot with Varying Prompts) and Fine-Tuned Model (M-Early: XLM-Roberta + CLIP + MLP) Across Three Random Subsets of the Development Set.

0.66, 0.69, and 0.68. The fine-tuning process allowed the model to learn the unique patterns and relationships between the visual and textual content of memes, resulting in a clear advantage over the zero-shot performance of GPT-4.

4.9.2 Analysis

GPT-4 Zero-Shot Performance

GPT-4's zero-shot performance shows incremental improvements as the prompt complexity increases, with the highest scores achieved using Prompt 4 (labels, definitions, and examples). The performance variability across the subsets was relatively minimal, suggesting that GPT-4 is capable of adapting to different memes with some consistency, especially when provided with detailed prompts. The standard deviation between subsets was not substantial, indicating that while GPT-4's reasoning might benefit from additional context, its predictions are relatively stable even in the zero-shot setting.

However, even with the most detailed prompt (Prompt 4), GPT-4's performance did not match that of the fine-tuned model. This suggests that while zero-shot learning is effective for general-purpose tasks, it faces limitations when applied to specialized tasks like persuasion detection in memes, where multimodal interactions are critical.

Comparison with Fine-Tuned Model

Despite GPT-4's broad language capabilities and adaptability, the fine-tuned XLM-Roberta + CLIP + MLP (M-Early) consistently outperformed it. The fine-tuned model's F1 scores across

all subsets were significantly higher, reflecting the benefits of task-specific training. Fine-tuning allowed the model to learn domain-specific patterns, particularly the nuanced interplay between text and images that is essential for detecting persuasive techniques in memes.

One key advantage of the fine-tuned model is its smaller size relative to GPT-4, making it more efficient in terms of training and inference costs. Fine-tuning a smaller model like `M-Early` requires far fewer resources than fine-tuning GPT-4, which, due to its large scale, demands significant computational power, financial investment, and contributes to a higher carbon footprint. Fine-tuning GPT-4 for a specific task like this would be prohibitively expensive for most research projects, making smaller models that can be fine-tuned effectively a more practical choice for specialized tasks.

The constraints of using the GPT-4 API also played a role in our decision to limit the zero-shot experiments to three random subsets of 100 memes each. Conducting the experiment on the entire development set would have been cost-prohibitive due to the API's paid nature. This highlights one of the challenges of using large models like GPT-4 for extensive experimentation without access to sufficient funding.

4.10 Chapter Summary

This chapter outlines our methodology for identifying persuasion techniques in multimodal meme content, combining textual and visual data. We focused on leveraging state-of-the-art neural architectures for detecting subtle persuasive strategies, addressing the challenge in Subtask 2 of SemEval 2024 Task 4 on Persuasion Detection in Memes.

We introduced two multi-label approaches: early fusion (`M-Early`) and cross-modal alignment (`M-Cross`). In the early fusion approach, image embeddings produced by CLIP and text embeddings generated by XLM-Roberta were concatenated and passed through an MLP. This allowed the model to process both image and text features at an early stage, enabling it to capture the interactions between the modalities and more effectively to detect persuasive elements within memes. This approach demonstrated the strongest performance in our experiments, underscoring the importance of joint multimodal processing.

In contrast, the cross-modal alignment approach (`M-Cross`), utilizing ResNet and BERT encodings processed through VisualBERT's cross-attention mechanism, did not achieve the same level of success. The alignment between the image and text modalities, while important for tasks such as visual question answering, proved less effective for the nuanced task of persuasion detection, as the multimodal interactions in memes often require a more integrated processing approach.

For comparitive purposes, we also experimented with single-modality models (`M-Image-Only` and `M-Text-Only`) (see Section 4.5.3 and Section 4.5.4), examining how well persuasion techniques could be detected when using either text or images alone. The results demonstrated that while images alone—handled through CLIP—could capture a significant amount of persuasive content, the highest performance was achieved when both text and images were processed together. Similarly, while text alone, processed through XLM-Roberta, performed reasonably well, the full persuasive message often relied on the context provided by the accompanying visual elements.

Additionally, we conducted zero-shot experiments using GPT-4 to evaluate its performance without prior training on the dataset. Four different prompt settings (see Section 4.6.1) were employed, ranging from a simple list of persuasion technique labels to a more comprehensive prompt that included definitions and examples. Although GPT-4 performed reasonably well in these zero-shot settings, its performance was consistently outperformed by the fine-tuned XLM-Roberta + CLIP + MLP model (`M-Early`), which was specifically trained on the task of persuasion detection. This comparison highlighted the benefits of task-specific fine-tuning, particularly for specialized tasks like meme-based persuasion detection, where multimodal interactions are critical.

The results and analysis section showcased the superiority of the fine-tuned models over GPT-4 in zero-shot settings. We demonstrated that while GPT-4 offers broad general-purpose capabilities, it is not as well-suited for highly specialized tasks that benefit from domain-specific training. Moreover, fine-tuning a smaller, task-specific model, like `M-Early`, is more resource-efficient compared to leveraging GPT-4, which would require significant computational resources, financial investment, and environmental costs to achieve similar results through fine-tuning.

In the next chapter, we will discuss the limitations of our work and outline potential directions for future work, building upon the insights gained from our experiments in multimodal persuasion detection.

Chapter 5

Conclusions and Future Work

This thesis aimed to address the challenge of detecting persuasion techniques in memes by leveraging both text and image modalities, contributing to the fields of NLP, multimodal learning, and disinformation tracking. By fine-tuning pre-trained models like BERT, XLM-RoBERTa, mBERT, CLIP, ResNet and VisualBERT, we tackled the SemEval 2024 Task 4 challenge, which involved multi-label classification across various persuasion techniques.

One of the key findings was that combining both modalities significantly enhanced the detection of persuasion techniques compared to single-modality approaches. This thesis successfully demonstrated how early fusion of text and image embeddings led to higher performance, whereas cross-modal alignment methods, like those in VisualBERT, underperformed. Additionally, the use of paraphrase-based data augmentation for the text-only task also contributed to improved detection by addressing the issue of class imbalance, particularly for underrepresented techniques. Augmenting text data with paraphrases allowed the models to learn better representations of rare persuasion techniques, which ultimately improved hierarchical F1 scores.

5.1 Contributions

This thesis makes three primary contributions. First, in the area of multi-label classification and data augmentation, we developed a system that fine-tuned pre-trained language models on textual

meme content to detect persuasion techniques, as detailed in Chapter 3. We applied paraphrase-based data augmentation to address the issue of class imbalance, particularly for underrepresented techniques. This resulted in a significant performance boost, showing that augmenting underrepresented classes helps achieve a more balanced classification output, as also demonstrated in our Para-Benef dataset experiments (Chapter 3). This contribution was published in Nayak and Kosseim (2024b) and Nayak and Kosseim (2024a).

Second, we explored multimodal persuasion detection by combining text and image embeddings through early fusion, which is described in Chapter 4. This approach outperformed both text-only and image-only models in the detection task. Interestingly, cross-modal alignment models like VisualBERT, which aim to align textual and visual elements in a more complex way, did not perform as well. This suggests that, for this particular task, early fusion strategies effectively captured multimodal interactions and that cross-modal alignment might have introduced additional complexity without significant benefit.

Lastly, we benchmarked zero-shot experiments with GPT-4 on a subset of the dataset (Chapter 4). Although GPT-4 showed potential in detecting persuasion techniques without fine-tuning, particularly when multimodal interactions were involved, our fine-tuned models outperformed it. This was particularly true for highly specialized tasks where pre-trained models with domain-specific tuning offered superior performance.

5.2 Limitations

Despite its contributions, the thesis has several limitations. One of the most prominent challenges was the dataset imbalance. While data augmentation helped alleviate the issue for some persuasion techniques, others remained underrepresented, which limited the model's ability to generalize across all techniques equally. The class imbalance was especially notable in persuasion techniques like "Repetition", which showed decreased performance even after augmentation, as detailed in Chapter 3.

Another limitation was the underperformance of cross-modal alignment models. While early fusion was highly effective, cross-modal approaches like VisualBERT did not yield the anticipated

improvements. This highlights the complexities of aligning visual and textual information effectively, especially when there is no clear correspondence between specific image regions and textual segments in memes. Moreover, we could have explored other embedding strategies to further enhance cross-modal alignment, as referenced in Chapter 4. Additionally, there is potential overfitting, particularly with more complex models like VisualBERT. The results suggest that while simpler methods like early fusion were robust, cross-modal models might have suffered from fitting too closely to the training data rather than generalizing well to unseen examples.

5.3 Future Work

Several promising avenues exist for future work based on this research. One major direction is the integration of hierarchical classification. Since persuasion techniques often follow a hierarchical structure, incorporating this into the model could lead to improved performance, as different techniques might share higher-level categories. This would provide a more structured and nuanced approach to the classification task, particularly in multi-label settings, as discussed in Chapter 3.

Another potential direction is refining the fusion process. While concatenating text and image embeddings proved effective, alternative methods such as attention-based mechanisms could offer a more sophisticated fusion of modalities. These methods could dynamically prioritize different aspects of the text and image, leading to better representation and interaction between modalities. This was touched upon in our multimodal experiments in Chapter 4, where we found that early fusion, though simple, was surprisingly effective.

Data augmentation techniques could also be extended to multimodal data. While our augmentation efforts focused on generating textual paraphrases, future research could involve augmenting both text and images. Techniques such as employing GPT-4 Vision or DALL-E to create alternate image variations would enrich the dataset and further address the class imbalance issue. This could result in a more balanced and diverse dataset for both text and image inputs.

Finally, exploring the impact of scaling models for both modalities represents an exciting direction. Larger language models, or vision language models, could significantly enhance the performance of multimodal detection systems. This would involve training on larger datasets and

potentially employing more computational resources, but the improvements in performance and generalization could be substantial.

References

Agarap, A. F. (2019). *Deep Learning using Rectified Linear Units (ReLU).* Retrieved from https://arxiv.org/abs/1803.08375

Alhuzali, H., & Ananiadou, S. (2021, April). SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2021)* (pp. 1573–1584). Online. Retrieved from https://aclanthology.org/2021.eacl-main.135

Barrón-Cedeño, A., Da San Martino, G., Jaradat, I., & Nakov, P. (2019, July). Proppy: A System to Unmask Propaganda in Online News. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-2019)*, *33*, 9847-9848. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/5061

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS-2020)* (Vol. 33, pp. 1877–1901). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cerri, R., Barros, R. C., & de Carvalho, A. C. (2014). Hierarchical Multi-label Classification Using Local Neural Networks. *Journal of Computer and System Sciences*, *80*(1), 39-56. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022000013000718

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., & Androutsopoulos, I. (2019, July). Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL-2019)* (pp. 6314–6322). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1636

Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D., Wang, Z., ... Liu, H. (2021). *MlTr: Multi-label Classification with Transformer.* Retrieved from https://arxiv.org/abs/2106.06195

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020, July). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)* (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.747

Costa, N. F., Hamilton, B., & Kosseim, L. (2023, July). CLaC at SemEval-2023 Task 3: Language Potluck RoBERTa Detects Online Persuasion Techniques in a Multilingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1613–1618). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.semeval-1.223

Da San Martino, G. (2019). *Detecting Propaganda in Online Media.* https://truthandtrustonline.com/wp-content/uploads/2019/10/paper_32.pdf. ([Accessed 15-08-2024])

Da San Martino, G., Barrón-Cedeño, A., & Nakov, P. (2019, November). Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. In *Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda (NLP4IF-2019)* (pp. 162–170). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-5024

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., & Nakov, P. (2020, December). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the 14th Workshop on Semantic Evaluation (SemEval-2020)* (pp. 1377–1414).

Barcelona (online): International Committee for Computational Linguistics. Retrieved from https://aclanthology.org/2020.semeval-1.186

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019, November). Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP-2019)* (pp. 5636–5646). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1565

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2019)* (pp. 4171–4186). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423

Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., & Da San Martino, G. (2024, June). SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 2009–2026). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.semeval-1.275

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* Retrieved from https://arxiv.org/abs/2010.11929

Feng, Z., Tang, J., Liu, J., Yin, W., Feng, S., Sun, Y., & Chen, L. (2021, August). Alpha at SemEval-2021 Task 6: Transformer Based Propaganda Classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 99–104). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.semeval-1.8

Goffredo, P., Chaves, M., Villata, S., & Cabrio, E. (2023, December). Argument-based Detection and Classification of Fallacies in Political Debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-2023)* (pp.

11101–11112). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.684

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition.* Retrieved from https://arxiv.org/abs/1512.03385

Howard, J., & Ruder, S. (2018, July). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-2018)* (pp. 328–339). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1031

Jacovi, A., Sar Shalom, O., & Goldberg, Y. (2018, November). Understanding Convolutional Neural Networks for Text Classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (EMNLP-2018)* (pp. 56–65). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-5408

Jiang, Y., Petrak, J., Song, X., Bontcheva, K., & Maynard, D. (2019, June). Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)* (pp. 840–844). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S19-2146

Jurkiewicz, D., Borchmann, Ł., Kosmala, I., & Graliński, F. (2020, December). ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)* (pp. 1415–1424). Barcelona (online): International Committee for Computational Linguistics. Retrieved from https://aclanthology.org/2020.semeval-1.187

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., ... Potthast, M. (2019, June). SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)* (pp. 829–839). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S19-2145

Kim, Y. (2014). Convolutional neural networks for sentence classification. *CoRR*, *abs/1408.5882*. Retrieved from http://arxiv.org/abs/1408.5882

Kudo, T., & Richardson, J. (2018, November). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-2018)* (pp. 66–71). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-2012

Lester, B., Al-Rfou, R., & Constant, N. (2021, November). The Power of Scale for Parameter-Efficient Prompt Tuning. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)* (pp. 3045–3059). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.243

Li, D., Wang, C., Zou, X., Wang, J., Chen, P., Wang, J., ... Lin, H. (2024, June). CoT-based Data Augmentation Strategy for Persuasion Techniques Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 1315–1321). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.semeval-1.190

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). *VisualBERT: A Simple and Performant Baseline for Vision and Language*. Retrieved from https://arxiv.org/abs/1908.03557

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... He, L. (2022, April). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2). Retrieved from https://doi.org/10.1145/3495162

Liao, Q., Lai, M., & Nakov, P. (2023, July). MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 83–87). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.semeval-1.10

Lin, N., Qin, G., Wang, G., Zhou, D., & Yang, A. (2023, July). An Effective Deployment of Contrastive Learning in Multi-label Text Classification. In *Findings of the Association for Computational Linguistics (ACL-2023)* (pp. 8730–8744). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.556

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*, 1 - 35. Retrieved from https://api.semanticscholar.org/CorpusID:236493269

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* Retrieved from https://arxiv.org/abs/1907.11692

Mapes, N., White, A., Medury, R., & Dua, S. (2019, November). Divisive Language and Propaganda Detection using Multi-head Attention Transformers with Deep Learning BERT-based Language Models for Binary Classification. In *Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda (NLP4IF-2019)* (pp. 103–106). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-5014

Milner, R. M., & Stephens, N. P. (2018, June). The World Made Meme: Public Conversations and Participatory Media. *International Journal of Communication*, *12*, 4. Retrieved from https://ijoc.org/index.php/ijoc/article/view/9696

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018, June). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)* (pp. 1–17). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S18-1001

Morishita, T., Morio, G., Ozaki, H., & Miyoshi, T. (2020, December). Hitachi at SemEval-2020 Task 3: Exploring the Representation Spaces of Transformers for Human Sense Word Similarity. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)* (pp. 286–291). Barcelona (online): International Committee for Computational Linguistics.

Retrieved from https://aclanthology.org/2020.semeval-1.36

Nakov, P., & Da San Martino, G. (2021). Fake News, Disinformation, Propaganda, Media Bias, and Flattening the Curve of the COVID-19 Infodemic. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (ACM-SIGKDD-2021)* (p. 4054–4055). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3447548.3470790

Nayak, K. S. R., & Kosseim, L. (2024a). *Analyzing Persuasive Strategies in Meme Texts: A Fusion of Language Models with Paraphrase Enrichment.* Retrieved from https://arxiv.org/abs/2407.01784

Nayak, K. S. R., & Kosseim, L. (2024b, June). CLaC at SemEval-2024 Task 4: Decoding Persuasion in Memes – An Ensemble of Language Models with Paraphrase Augmentation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 175–180). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.semeval-1.27

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., ... Odena, A. (2021). *Show Your Work: Scratchpads for Intermediate Computation with Language Models.* Retrieved from https://arxiv.org/abs/2112.00114

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ... Zoph, B. (2024). *GPT-4 Technical Report.* Retrieved from https://arxiv.org/abs/2303.08774

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks.* Retrieved from https://arxiv.org/abs/1511.08458

Piskorski, J., N. Stefanovitch, V.-A. B., Faggiani, N., Linge, J., Kharazi, S., Nikolaidis, N., ... Nakov, P. (2023). *News Categorization, Framing and Persuasion Techniques: Annotation Guidelines.* European Commission Joint Research Centre. Retrieved from https://knowledge4policy.ec.europa.eu/sites/default/files/JRC132862_technical_report_annotation_guidelines_final_with_affiliations_1.pdf ([Accessed 15-08-2024])

Piskorski, J., Stefanovitch, N., Da San Martino, G., & Nakov, P. (2023, July). 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)* (pp. 2343–2361). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.semeval-1.317

Radford, A. (2018). Improving language understanding by generative pre-training.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, 18–24 Jul). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML-2021)* (Vol. 139, pp. 8748–8763). Proceedings of Machine Learning Research (PMLR). Retrieved from https://proceedings.mlr.press/v139/radford21a.html

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.* Retrieved from https://api.semanticscholar.org/CorpusID:160025533

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier Chains for Multi-label Classification. In *Machine Learning and Knowledge Discovery in Databases* (pp. 254–269). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* Retrieved from https://arxiv.org/abs/1506.01497

Schmidhuber, J. (2015, January). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. Retrieved from http://dx.doi.org/10.1016/j.neunet.2014.09.003

Shifman, L., & Handloff, J. R. (2015, December). Memes in digital culture. *The Communication Review*, *18*(4), 315–318. Retrieved from https://www.tandfonline.com/doi/full/10.1080/10714421.2015.1100480

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017, sep). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorer Newsletter*, *19*(1), 22–36. Retrieved from https://doi.org/10.1145/3137597.3137600

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition.* Retrieved from https://arxiv.org/abs/1409.1556

Tan, H., & Bansal, M. (2019, November). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP-2019)* (pp. 5100–5111). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1514

Tian, J., Gui, M., Li, C., Yan, M., & Xiao, W. (2021, August). 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)* (pp. 1082–1087). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.semeval-1.150

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). *Mining Multi-label Data.* Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-09823-4_34

Tsoumakas, K., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems (NeurIPS-2017)* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2023). *Self-Consistency Improves Chain of Thought Reasoning in Language Models.* Retrieved from https://arxiv.org/abs/2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* Retrieved from https://arxiv.org/abs/2201.11903

Wendling, M. (2018). *Alt-right: From 4chan to the White House*. Pluto Press.

Wu, B., Razuvayevskaya, O., Heppell, F., Leite, J. A., Scarton, C., Bontcheva, K., & Song,

X. (2023, July). SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1995–2008). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.semeval-1.275

Wunderle, J., Schubert, J., Cacciatore, A., Zehe, A., Pfister, J., & Hotho, A. (2024, June). OtterlyObsessedWithSemantics at SemEval-2024 Task 4: Developing a Hierarchical Multi-Label Classification Head for Large Language Models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 602–612). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.semeval-1.90

Yang, Z., Yatskar, M., Huang, A., Dyer, C., Devlin, J., & Lee, K. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2016)* (pp. 1480–1489). Association for Computational Linguistics.

Yoosuf, S., & Yang, Y. (2019, November). Fine-Grained Propaganda Detection with Fine-Tuned BERT. In *Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda (NLP4IF-2019)* (pp. 87–91). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-5011

Yu, C., Shen, Y., Mao, Y., & Cai, L. (2022). Constrained Sequence-to-Tree Generation for Hierarchical Text Classification. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM-SIGIR-2022)*. Retrieved from https://api.semanticscholar.org/CorpusID:247939715

Yu, S., Martino, G. D. S., Mohtarami, M., Glass, J., & Nakov, P. (2021). *Interpretable Propaganda Detection in News Articles.* Retrieved from https://arxiv.org/abs/2108.12802

Yu, S., Martino, G. D. S., & Nakov, P. (2019). Experiments in Detecting Persuasion Techniques in the News. *CoRR*, *abs/1911.06815*. Retrieved from http://arxiv.org/abs/1911.06815

Zhou, J., Ma, C., Long, D., Xu, G., Ding, N., Zhang, H., ... Liu, G. (2020, July). Hierarchy-Aware Global Model for Hierarchical Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)* (pp. 1106–1117). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.104

Appendix A

Appendix: Persuasion Techniques and their definitions

Source of the definitions: https://propaganda.math.unipd.it/semeval2024task4/definitions22.html

(1) **Presenting Irrelevant Data (Red Herring)**: Introducing irrelevant material to the issue being discussed to divert attention away from the points made.

(2) **Misrepresentation of Someone's Position (Straw Man)**: Substituting an opponent's proposition with a similar one that is easier to refute.

(3) **Whataboutism**: Attempting to discredit an opponent's position by charging them with hypocrisy without addressing the original argument.

(4) **Causal Oversimplification**: Assuming a single cause or reason for an issue when multiple factors are involved.

(5) **Obfuscation, Intentional Vagueness, Confusion**: Using unclear language that allows for multiple interpretations, weakening the support for a conclusion.

(6) **Appeal to Authority**: Stating that a claim is true simply because a valid authority or expert on the issue said so, without offering other supporting evidence.

(7) **Black-and-white Fallacy (Dictatorship)**: Presenting two alternatives as the only possibilities when others exist.

(8) **Name Calling or Labeling**: Labeling the subject in a way that evokes fear, hatred, or admiration, affecting the audience's perception.

(9) **Loaded Language**: Using emotionally charged words or phrases to influence an audience's perception.

(10) **Exaggeration or Minimisation**: Representing something in an excessive or diminished manner to alter its perceived importance.

(11) **Flag-waving**: Invoking strong feelings of patriotism or loyalty to a group to justify or promote an action or idea.

(12) **Doubt**: Questioning the credibility of a person or thing without substantial evidence.

(13) **Appeal to Fear/Prejudice**: Seeking to build support by inciting fear or preconceived judgments against an alternative.

(14) **Slogans**: Using brief, striking phrases that appeal emotionally to the audience.

(15) **Thought-terminating Cliché**: Using phrases that discourage further thought or discussion about a complex topic.

(16) **Bandwagon**: Persuading an audience to take action because "everyone else is doing it."

(17) **Reductio ad Hitlerum**: Discrediting an idea by associating it with disliked or despised groups or individuals.

(18) **Repetition**: Repeating the same message frequently to ensure it is accepted by the audience.

(19) **Smears**: Damaging someone's reputation by spreading negative propaganda about them.

(20) **Glittering Generalities**: Using positive, value-laden words or symbols to create a favorable image.

(21) **Transfer**: Associating positive or negative qualities of one thing with another to alter the audience's perception.

(22) **Appeal to (Strong) Emotions**: Using images or language that provoke strong emotional responses to influence an audience.

Appendix B

Appendix: Zero-shot Experiment Prompts and API Query Format

This appendix contains the format for querying the GPT-4 API[1] to predict the most likely persuasion techniques as well as our different prompt settings used in our zero-shot experiments with GPT-4 for the task of persuasion technique detection (see Section 4.6).

The prompts vary in complexity, from providing only the list of techniques to including definitions and examples for each technique.

B.1 API Query Format

The following Python code shows how we queried the GPT-4 API to identify persuasion techniques within memes. The query uses both text and image (encoded as base64), and the response is expected in a specific JSON format.

[1] https://platform.openai.com/docs/models/gpt-4o

```
def query_api(text, base64_image):
   response = client.chat.completions.create(
      model="gpt-4o",
      response_format={"type": "json_object"},
      temperature=0.9,
      messages=[
         {
            "role": "user",
            "content": [
               {"type": "text", "text": f'''
                  {prompt-n}
                  As additional information, here is the text that is
                     embedded in the image: {text}
                  Identify the techniques conveyed by the image and
                     provide the results
                  as a JSON object with the key "techniques" and a list
                     of the identified
                  techniques as the value.
                  The output should be in the following format:
                  ```json
                  {{
                     "techniques": ["technique1", "technique2", ...]
                  }}
                  ```
                  Note: Only include the techniques that are identified
                     in the image
                  (most probable).
               '''},
               {
                  "type": "image_url",
                  "image_url": {
                     "url": f"{base64_image}",
                  },
               },
            ],
         }
      ],
      max_tokens=100,
)
   return response.choices[0].message.content
```

B.2 Prompt 1: Persuasion Techniques

Prompt 1 includes only the list of persuasion techniques.

Given the following list of persuasion techniques and their definitions:

```
techniques_list = [
"Repetition",
"Obfuscation, Intentional vagueness, Confusion",
"Causal Oversimplification",
"Black-and-white Fallacy/Dictatorship",
"Thought-terminating cliche",
"Misrepresentation of Someone's Position (Straw Man)",
"Presenting Irrelevant Data (Red Herring)",
"Whataboutism",
"Slogans",
"Bandwagon",
"Appeal to authority",
"Flag-waving",
"Appeal to fear/prejudice",
"Glittering generalities (Virtue)",
"Doubt",
"Name calling/Labeling",
"Smears",
"Reductio ad hitlerum",
"Transfer",
"Exaggeration/Minimisation",
"Loaded Language",
"Appeal to (Strong) Emotions"]
```

B.3 Prompt 2: Persuasion Techniques with Definitions

Prompt 2 includes:

- List of persuasion techniques.
- Definitions of persuasion techniques.

Given the following list of persuasion techniques and their definitions:

```
techniques_list = [
"Repetition",
"Obfuscation, Intentional vagueness, Confusion",
"Causal Oversimplification",
"Black-and-white Fallacy/Dictatorship",
"Thought-terminating cliche",
"Misrepresentation of Someone's Position (Straw Man)",
"Presenting Irrelevant Data (Red Herring)",
"Whataboutism",
"Slogans",
"Bandwagon",
"Appeal to authority",
"Flag-waving",
"Appeal to fear/prejudice",
"Glittering generalities (Virtue)",
"Doubt",
"Name calling/Labeling",
"Smears",
"Reductio ad hitlerum",
"Transfer",
"Exaggeration/Minimisation",
"Loaded Language",
"Appeal to (Strong) Emotions"]
```

Definitions

Presenting Irrelevant Data (Red Herring): Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.

Misrepresentation of Someone's Position (Straw Man): When an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition.

Whataboutism: A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

Causal Oversimplification: Assuming a single cause or reason when there are actually multiple causes for an issue. It includes transferring blame to one person or group of people without investigating the complexities of the issue.

Obfuscation, Intentional vagueness, Confusion: Using words which are deliberately not clear so that the audience may have its own interpretations.

Appeal to authority: Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered.

Black-and-white Fallacy/Dictatorship: Presenting two alternative options as the only possibilities, when in fact more possibilities exist.

Name calling/Labeling: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises.

Loaded Language: Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.

Exaggeration or Minimisation: Either representing something in an excessive manner or making something seem less important or smaller than it really is.

Flag-waving: Playing on strong national feeling (or to any group; e.g., race, gender, political preference) to justify or promote an action or idea.

Doubt: Questioning the credibility of someone or something.

Appeal to fear/prejudice: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative.

Slogans: A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

Thought-terminating cliche: Words or phrases that discourage critical thought and meaningful discussion about a given topic.

Bandwagon: Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action."

Reductio ad hitlerum: Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience.

Repetition: Repeating the same message over and over again so that the audience will eventually accept it.

Smears: A smear is an effort to damage or call into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups.

Glittering Generalities: These are words or symbols in the value system of the target audience that produce a positive image when attached to a person or issue. Examples: Peace, hope, happiness, security, wise leadership, etc.

Transfer: Also known as association, this is a technique of projecting positive or negative qualities (praise or blame) of a person, entity, object, or value onto another to make the second more acceptable or to discredit it. It evokes an emotional response, which stimulates the target to identify with recognized authorities. Often highly visual, this technique often utilizes symbols (for example, the swastikas used in Nazi Germany, originally a symbol for health and prosperity) superimposed over other visual images.

Appeal to (strong) Emotions: Using images with strong positive/negative emotional implications to influence an audience.

B.4 Prompt 3: Persuasion Techniques with Examples

Prompt 3 includes:

- List of persuasion techniques.
- Examples of persuasion techniques.

Given the following list of persuasion techniques and examples for each of them:

```
techniques_list = [
"Repetition",
"Obfuscation, Intentional vagueness, Confusion",
"Causal Oversimplification",
"Black-and-white Fallacy/Dictatorship",
"Thought-terminating cliche",
"Misrepresentation of Someone's Position (Straw Man)",
"Presenting Irrelevant Data (Red Herring)",
"Whataboutism",
"Slogans",
"Bandwagon",
"Appeal to authority",
"Flag-waving",
"Appeal to fear/prejudice",
"Glittering generalities (Virtue)",
"Doubt",
"Name calling/Labeling",
"Smears",
"Reductio ad hitlerum",
"Transfer",
"Exaggeration/Minimisation",
"Loaded Language",
"Appeal to (Strong) Emotions"]
```

Examples

Presenting Irrelevant Data (Red Herring):

Example 1: In politics, defending one's own policies regarding public safety - "I have worked hard to help eliminate criminal activity. What we need is economic growth that can only come from the hands of leadership."

Example 2: You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?

Misrepresentation of Someone's Position (Straw Man):

Example 1: Zebedee: What is your view on the Christian God?

Mike: I don't believe in any gods, including the Christian one.

Zebedee: So you think that we are here by accident, and all this design in nature is pure chance, and the universe just created itself?

Mike: You got all that from me stating that I just don't believe in any gods?

Example 2: Putin: When I'm done with Ukraine I'm coming to South Africa, I want you to tell me why you call sausages Russians.

Whataboutism:

Example 1: A nation deflects criticism of its recent human rights violations by pointing to the history of slavery in the United States.

Example 2: Qatar spending profusely on Neymar, not fighting terrorism.

Causal Oversimplification:

Example 1: President Trump has been in office for a month and gas prices have been skyrocketing. The rise in gas prices is because of President Trump.

Example 2: The reason New Orleans was hit so hard with the hurricane was because of all the immoral people who live there.

Example 3: If France had not declared war on Germany then World War II would have never happened.

Obfuscation, Intentional vagueness, Confusion:

Example 1: It is a good idea to listen to victims of theft. Therefore, if the victims say to have the thief shot, then you should do that.

Example 2: It's beginning to look a lot like I told you so.

Appeal to authority:

Example 1: Richard Dawkins, an evolutionary biologist and perhaps the foremost expert in the field, says that evolution is true. Therefore, it's true.

Example 2: According to Serena Williams, our foreign policy is the best on Earth. So we are in the right direction.

Black-and-white Fallacy/Dictatorship:

Example 1: You must be a Republican or Democrat. You are not a Democrat. Therefore, you must be a Republican.

Example 2: I thought you were a good person, but you weren't at church today.

Example 3: There is no alternative to war.

Name calling/Labeling:

Example 1: Republican congressweasels.

Example 2: Bush the Lesser (note that lesser does not refer to "the second", but it is pejorative).

Loaded Language:

Example 1: [...] a lone lawmaker's childish shouting.

Example 2: How stupid and petty things have become in Washington.

Exaggeration or Minimisation:

Example 1: Democrats bolted as soon as Trump's speech ended in an apparent effort to signal they can't even stomach being in the same room as the president.

Example 2: We're going to have unbelievable intelligence.

Example 3: I was not fighting with her; we were just playing.

Flag-waving:

Example 1: Patriotism means no questions.

Example 2: Entering this war will make us have a better future in our country.

Doubt:

Example 1: A candidate talks about his opponent and says: Is he ready to be the Mayor?

Example 2: OH YOU DON'T BELIEVE AMERICA WOULD EVER USE THE NDAA TO INDEFINITELY DETAIN ITS CITIZENS IN FEMA CAMPS? PLEASE TELL ME MORE ABOUT HOW JAPANESE-AMERICAN INTERNMENT CAMPS DURING WWII ARE JUST A CONSPIRACY THEORY.

Appeal to fear/prejudice:

Example 1: Either we go to war or we will perish.

Example 2: We must stop those refugees as they are terrorists.

Slogans:

Example 1: The more women at war . . . the sooner we win.

Example 2: Make America great again!

Thought-terminating cliche:

Example 1: It is what it is.

Example 2: It's just common sense.

Example 3: Nothing is permanent except change.

Bandwagon:

Example 1: Would you vote for Clinton as president? 57% say yes.

Example 2: 90% of citizens support our initiative. You should.

Reductio ad hitlerum:

Example 1: Do you know who else was doing that? Hitler!

Example 2: Only one kind of person can think in that way: a communist.

Repetition:

Example 1: Losers, defeated, losers defeated!

Example 2: I voted for him once...and I will vote for him again.

Smears:

Example 1: A MORE DESPICABLE COWARD OF A POLITICIAN YOU WILL NOT FIND A TRAITOR AND A THIEF!

Example 2: RUSSIA'S OFERTON UKRAINE'S OPERATION.

Glittering Generalities:

Example 1: I PLEDGE TO VOTE FOR THIS MAN AGAINND I TRUST HIM MORE THAN

ANY MEDIA COMPANY ON THE PLANET.

Example 2: GOOD MORNING FELLOW PATRIOTS!

Transfer:

Example 1: HILLARY WILL BE A GREAT PRESIDENT. SHE'LL OPPRESS WORKERS AND ADVANCE IMPERIALISM JUST LIKE THE BEST OF THEM.

Example 2: Listen here Jack. Cry harder.

Appeal to (strong) Emotions:

Example 1: PEACEFUL PROTESTORS, DETAINED AND ARRESTED ACROSS RUSSIA, MUST BE RELEASED IMMEDIATELY.

Example 2: May your Thanksgiving be filled with the exact opposite spirit of U.S. governors trying to keep out refugees.

B.5 Prompt 4: Persuasion Techniques, Definitions, and Examples

Prompt 4 includes:

- List of persuasion techniques.
- Definitions of persuasion techniques.
- Examples of persuasion techniques.

Given the following list of persuasion techniques, definitions, and examples:

```
techniques_list = [
"Repetition",
"Obfuscation, Intentional vagueness, Confusion",
"Causal Oversimplification",
"Black-and-white Fallacy/Dictatorship",
"Thought-terminating cliche",
"Misrepresentation of Someone's Position (Straw Man)",
"Presenting Irrelevant Data (Red Herring)",
"Whataboutism",
"Slogans",
"Bandwagon",
"Appeal to authority",
"Flag-waving",
"Appeal to fear/prejudice",
"Glittering generalities (Virtue)",
"Doubt",
"Name calling/Labeling",
"Smears",
"Reductio ad hitlerum",
"Transfer",
"Exaggeration/Minimisation",
"Loaded Language",
"Appeal to (Strong) Emotions"]
```

Definitions and Examples

Presenting Irrelevant Data (Red Herring)

Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.

Example 1: In politics, defending one's own policies regarding public safety - "I have worked hard to help eliminate criminal activity. What we need is economic growth that can only come from the hands of leadership."

Example 2: You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?

Misrepresentation of Someone's Position (Straw Man)

When an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition.

Example 1: Zebedee: What is your view on the Christian God? Mike: I don't believe in any gods, including the Christian one. Zebedee: So you think that we are here by accident, and all this design in nature is pure chance, and the universe just created itself? Mike: You got all that from me stating that I just don't believe in any gods?

Example 2: Putin: When I'm done with Ukraine I'm coming to South Africa, I want you to tell me why you call sausages Russians.

Whataboutism

A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

Example 1: A nation deflects criticism of its recent human rights violations by pointing to the history of slavery in the United States.

Example 2: Qatar spending profusely on Neymar, not fighting terrorism.

Causal Oversimplification

Assuming a single cause or reason when there are actually multiple causes for an issue. It includes transferring blame to one person or group of people without investigating the complexities of the

issue.

Example 1: President Trump has been in office for a month and gas prices have been skyrocketing. The rise in gas prices is because of President Trump.

Example 2: The reason New Orleans was hit so hard with the hurricane was because of all the immoral people who live there.

Example 3: If France had not declared war on Germany then World War II would have never happened.

Obfuscation, Intentional vagueness, Confusion

Using words which are deliberately not clear so that the audience may have its own interpretations.

Example 1: It is a good idea to listen to victims of theft. Therefore if the victims say to have the thief shot, then you should do that.

Example 2: It's beginning to look a lot like I told you so.

Appeal to authority

Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered.

Example 1: Richard Dawkins, an evolutionary biologist and perhaps the foremost expert in the field, says that evolution is true. Therefore, it's true.

Example 2: According to Serena Williams, our foreign policy is the best on Earth. So we are in the right direction.

Black-and-white Fallacy/Dictatorship

Presenting two alternative options as the only possibilities, when in fact more possibilities exist.

Example 1: You must be a Republican or Democrat. You are not a Democrat. Therefore, you must be a Republican.

Example 2: I thought you were a good person, but you weren't at church today.

Example 3: There is no alternative to war.

Name calling/Labeling

Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises.

Example 1: Republican congressweasels.

Example 2: Bush the Lesser (note that lesser does not refer to ”the second”, but it is pejorative).

Loaded Language

Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.

Example 1: [...] a lone lawmaker’s childish shouting.

Example 2: How stupid and petty things have become in Washington.

Exaggeration or Minimisation

Either representing something in an excessive manner or making something seem less important or smaller than it really is.

Example 1: Democrats bolted as soon as Trump’s speech ended in an apparent effort to signal they can’t even stomach being in the same room as the president.

Example 2: We’re going to have unbelievable intelligence.

Example 3: I was not fighting with her; we were just playing.

Flag-waving

Playing on strong national feeling (or to any group; e.g., race, gender, political preference) to justify or promote an action or idea.

Example 1: Patriotism means no questions.

Example 2: Entering this war will make us have a better future in our country.

Doubt

Questioning the credibility of someone or something.

Example 1: A candidate talks about his opponent and says: Is he ready to be the Mayor?

Example 2: OH YOU DON’T BELIEVE AMERICA WOULD EVER USE THE NDAA TO INDEFINITELY DETAIN ITS CITIZENS IN FEMA CAMPS? PLEASE TELL ME MORE ABOUT HOW JAPANESE-AMERICAN INTERNMENT CAMPS DURING WWII ARE JUST A CONSPIRACY THEORY.

Appeal to fear/prejudice

Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative.

Example 1: Either we go to war or we will perish.

Example 2: We must stop those refugees as they are terrorists.

Slogans

A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

Example 1: The more women at war . . . the sooner we win.

Example 2: Make America great again!

Thought-terminating cliche

Words or phrases that discourage critical thought and meaningful discussion about a given topic.

Example 1: It is what it is.

Example 2: It's just common sense.

Example 3: Nothing is permanent except change.

Bandwagon

Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action".

Example 1: Would you vote for Clinton as president? 57% say yes.

Example 2: 90% of citizens support our initiative. You should.

Reductio ad hitlerum

Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience.

Example 1: Do you know who else was doing that? Hitler!

Example 2: Only one kind of person can think in that way: a communist.

Repetition

Repeating the same message over and over again so that the audience will eventually accept it.

Example 1: Losers, defeated, losers defeated!

Example 2: I voted for him once...and I will vote for him again.

Smears

A smear is an effort to damage or call into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups.

Example 1: A MORE DESPICABLE COWARD OF A POLITICIAN YOU WILL NOT FIND A

TRAITOR AND A THIEF!

Example 2: RUSSIA'S OFERTON UKRAINE'S OPERATION.

Glittering Generalities

These are words or symbols in the value system of the target audience that produce a positive image when attached to a person or issue. Examples: Peace, hope, happiness, security, wise leadership, etc.

Example 1: I PLEDGE TO VOTE FOR THIS MAN AGAINND I TRUST HIM MORE THAN ANY MEDIA COMPANY ON THE PLANET.

Example 2: GOOD MORNING FELLOW PATRIOTS!

Transfer

Also known as association, this is a technique of projecting positive or negative qualities (praise or blame) of a person, entity, object, or value onto another to make the second more acceptable or to discredit it. It evokes an emotional response, which stimulates the target to identify with recognized authorities. Often highly visual, this technique often utilizes symbols (for example, the swastikas used in Nazi Germany, originally a symbol for health and prosperity) superimposed over other visual images.

Example 1: HILLARY WILL BE A GREAT PRESIDENT. SHE'LL OPPRESS WORKERS AND ADVANCE IMPERIALISM JUST LIKE THE BEST OF THEM.

Example 2: Listen here Jack. Cry harder.

Appeal to (strong) Emotions

Using images with strong positive/negative emotional implications to influence an audience.

Example 1: PEACEFUL PROTESTORS, DETAINED AND ARRESTED ACROSS RUSSIA, MUST BE RELEASED IMMEDIATELY.

Example 2: May your Thanksgiving be filled with the exact opposite spirit of U.S. governors trying to keep out refugees.