Dissecting Persuasive Text: Models for Technique Classification and Span Detection

CSI4900
Honours Project

Matteo Torlone 300248556
Alone Petrova 300074852

University of Ottawa
April 24, 2025

Abstract

Propaganda in online media operates through subtle rhetorical strategies that shape beliefs and emotions without relying on factual arguments. As social platforms accelerate the spread of persuasive content, ranging from emotionally charged slogans to misleading causal claims, automated detection of these tactics becomes critical for preserving informed discourse. In this project, we undertake SemEval-2021 Task 6's two complementary challenges: first, identifying which of twenty fine-grained persuasion techniques appear in a short text snippet; and second, pinpointing the exact spans where each technique occurs. Our investigation shows that modern Transformer architectures can learn the complex signals of loaded language, name-calling, appeals to fear, and other tactics when framed as multi-label classification, yet they struggle to localize span boundaries under scarce annotations. Performance on rare techniques remains limited by data imbalance and boundary ambiguity, which points to the need for synthetic span augmentation, unified span classification models, and span-aware representation learning. By mapping the conceptual landscape of persuasion detection, drawing on both SemEval-vintage pipelines and recent advances in question-answering and type-aware modeling, we outline a research agenda to close the gap between text-level identification and precise span-level explanation.

1. Introduction

In the age of digital communication and internet news, persuasive language can heavily influence public opinion. From politics to business, persuasion skills are employed in an attempt to sway readers and viewers. Propaganda refers to a specific type of strategic communication aimed to influence the audience's beliefs, emotions, and even actions. This is often done through unbalanced or emotionally charged content (Jowett, O'Donnel 2018). Propaganda often resorts to persuasion tactics such as linguistic or rhetorical devices that operate on emotion, authority, or group affiliation and thus shape public opinion without relying on factual truth.
The influence of propaganda has grown exponentially. Social media platforms, in particular, have become fertile soil for the rapid spread of persuasive and manipulative information. Studies have shown that false or emotionally manipulated information spreads much more quickly and widely than factual news (Vosoughi, Roy, & Aral, 2018). The use of subtle techniques, such as appeals to fear, name-calling, and whataboutism makes such content even more difficult to recognize, especially for the general reader.
Given the sophistication and effect of modern propaganda, there is an urgent need for digital systems to detect persuasive content across various media. Manual content moderation is not scalable or reproducible. Therefore, most Natural Language Processing researchers have begun developing models to detect the use of propaganda and rhetorical devices. Transformer models such as BERT and RoBERTa have demonstrated promising results in the detection of context-specific persuasive language patterns (Da San Martino et al., 2020).
Social media websites and content moderation software currently employ a combination of rule-based filtering, keyword detection, and user reporting to combat misinformation and damaging persuasion. While these approaches have been somewhat successful at detecting false content, they are not successful at detecting more subtle forms of propaganda. The

complexity of persuasive communication frequently renders binary classification models ineffective. Moreover, existing approaches are typically not interpretable and do not provide clear explanations of how or why a given piece of content is considered persuasive.
These limitations are illustrated through real-world misclassifications. For instance, Facebook's AI moderation tool mistakenly labeled the Auschwitz Museum posts as violating community standards (Futurism, 2024; Sircar, 2024). Similarly, a YouTube livestream featuring chess grandmaster Hikaru Nakamura was abruptly taken down from the platform after the site's automated system flagged a discussion of the "King's Indian Defense" as "harmful and dangerous" content despite the fact it is a normal chess term with no malicious intent (Wired, 2021). These instances illustrate the risk of over-reliance on rigid or context-insensitive moderation pipelines and emphasize the need for more context-sensitive and explainable tools that can accurately examine persuasive language.
Numerous global efforts have been proposed to advance even more in this field. An example of those is SemEval (Semantic Evaluation), an established series of competitions in NLP designed to judge semantic analysis systems on a set of challenging tasks. SemEval-2021 Task 6, Detection of Persuasion Techniques in Texts and Images, set a goal of further pushing the borders of computational argumentation and media analysis by focusing on fine-grained detection of persuasive tactics (Dimitrov et al., 2021). As part of our honors project, our team decided to take on two of the task's subtasks:

- Subtask 1: A multi-label classification task focused on identifying which persuasion techniques were present in a given text fragment.

- Subtask 2: A span-level identification task requiring systems to locate the specific segments of text where each technique appeared, along with their corresponding labels.

Our primary goal was to design and evaluate machine learning models capable of tackling these tasks. We also aimed to compare their performance with SemEval's participant submissions. And finally, reflect on the implications for real-world applications in content moderation and media analysis.

2.1 Background information

Taxonomy of persuasion techniques

Persuasion techniques are linguistic strategies aiming at influencing public opinion, emotions, and even action. They often appear in political debates, news, and social media (Bassi, Fomsgaard, & Pereira-Fariña, 2024). These techniques have a tendency to appeal to various emotions and social processes.

The SemEval-2021 Task 6 focuses on the identification of 20 propaganda techniques used in memes in both textual and visual components (Dimitrov et al., 2021). These techniques were collected from various works such as Da San Martino et al. (2019b) and Shah (2005), which are aimed at manipulating the audience through different forms of persuasive tactics.

These 20 techniques can be divided into a few categories based on their underlying strategies: emotional appeals, logical manipulations, and social influence. The following table describes the 20 techniques that apply to both text and images (Dimitrov et al., 2021).

Technique	Description
Loaded Language	Using emotionally charged words to influence perceptions.
Name-Calling or Labeling	Assigning negative or positive labels to individuals or ideas.
Doubt	Questioning the credibility of individuals or ideas.
Exaggeration or Minimization	Overstating or understating facts to manipulate perceptions.
Appeal to Fear or Prejudices	Inducing fear or reinforcing prejudices to persuade.
Slogans	Brief, emotionally charged phrases meant to persuade.
Whataboutism	Diverting attention to accusations of hypocrisy rather than addressing the issue.
Flag-Waving	Leveraging national or group sentiment to justify actions.
Straw Man	Misrepresenting an argument to easily refute it.
Causal Oversimplification	Presenting one cause for a complex issue.
Appeal to Authority	Citing authorities or experts without further evidence.
Thought-Terminating Cliché	Using dismissive phrases to avoid discussion.

Black-and-White Fallacy	Presenting only two extreme options.
Reductio ad Hitlerum	Associating ideas with negative historical figures to discredit them.
Repetition	Repeating the message to make it more believable.
Obfuscation	Using vague language to confuse or leave room for interpretation.
Red Herring	Introducing irrelevant material to distract from the issue.
Bandwagon	Encouraging people to follow others for social conformity.
Smears	Damaging someone's reputation through negative statements.
Glittering Generalities	Using emotionally appealing values to positively sway opinions.

Computational Approaches to Detecting Persuasion

The detection of persuasive language has become an increasingly important topic in Natural Language Processing (NLP). Various computational models were developed to identify persuasion techniques. In the early stages, rule-based systems were proposed to identify persuasive language (Bassi, Fomsgaard, & Pereira-Fariña, 2024). These systems usually involved predefined lists of words and phrases that were heuristically chosen based on prior research. While such approaches were considered relatively simple and interpretable, they often struggled with scalability and flexibility, especially when dealing with large datasets (Pecan Team, 2023).

Machine learning models such as Support Vector Machines (SVM) were also used to classify persuasive content based on features extracted from the text (Bassi, Fomsgaard, & Pereira-Fariña, 2024). However, SVM-based approaches require extensive feature engineering, where researchers manually identify features, such as word n-grams, sentence structure, and sentiment to train the model (Abd Kadir & Sauffiyan, 2019). Despite improved performance in contrast to rule-based systems, SVM's struggled to handle complex patterns and understand the context-based nuances of persuasive communication

In recent years, transformer-based models, such as BERT and RoBERTa set new benchmarks for text classification tasks, including the detection of persuasive language (Liu et al., 2019). These models are based on self-attention mechanisms that allow them to capture contextual

relationships between words in a sentence, enabling them to understand the meaning of a text at a much deeper level than traditional models.
Despite the success of transformer models, several key challenges persist in detecting persuasive language: ambiguity, overlap of the persuasive techniques, and fine-grained detection. Many persuasive techniques are subtle and heavily rely on context-dependent cues. In addition, persuasive techniques often overlap or co-occur in the same text, which will be showcased in our Dataset section. Another challenge is the fine-grained nature of the task. Detecting the exact boundaries where a persuasion technique occurs in a text requires high levels of precision.

Related Works and Literature Review

Current research heavily leverages pre-trained language models such as BERT and RoBERTa, that are fine-tuned on annotated datasets containing propaganda or persuasion techniques to improve capabilities for detection from text and multimodal content(Abdullah, Altiti, & Obiedat, 2019).
One of the most influential efforts in this space is by Da San Martino et al. (2020), who introduced a fine-grained propaganda detection dataset and taxonomy, which helped standardize the task and pinpoint essential challenges such as multi-label nature of persuasion, the vagueness of categories, and the necessity of contextual analysis. Complementing this, Da San Martino et al. (2020b) also developed PRTA (Propaganda Techniques Analysis), a system that allows for comprehensive analysis of propaganda in news. It introduced an interface and computational capabilities to annotate, visualize, and analyze propaganda across various textual elements, further driving the development of explainable AI systems in this domain.
Several studies have expanded the scope of persuasion detection beyond text, recognizing that modern propaganda often involves a multimodal approach, combining text with visual cues. Models such as VisualBERT and CLIP have been applied to these tasks, aiming to detect persuasive framing or emotional appeals in image-text combinations (Li, Yatskar, Yin, Hsieh, & Chang, 2019) (Ghadery, Sileo, & Moens, 2021).
Very recent research has suggested cutting-edge techniques for propaganda and persuasive content identification in both text and multimodal environments. Muthukumar et al. (2024) proposed a deepfake detection system using an Adaptive Gaining Sharing Knowledge (AGSK) algorithm with DenseNet121, which achieved 98% accuracy in identifying manipulated and propagandist images. Wang and Chen (2025, IEEE) proposed a hybrid detection platform for identifying political propaganda in social media template-based images by combining object detection, text analysis, and pixel-level features. Cui et al. (2024) introduced a state-of-the-art multimodal model—APCL—that employs anti-persuasion prompts and contrastive learning to recognize logical fallacies in meme-based propaganda more effectively. Also in 2025, Ahmad et al. presented a Hierarchical Graph-based Integration Network (H-GIN) designed for textual propaganda detection, which integrates semantic, syntactic, and sequential features using a bi-layer graph model, outperforming several baseline models on benchmark datasets.
Several competitive systems participated in the SemEval 2021 task, and transformer models performed best in the text-based subtasks. Team MinD emerged as the winner of Subtask 1 through the application of five transformers—BERT, RoBERTa, XLNet, DeBERTa, and

ALBERT—combined with effective post-processing (Tian et al., 2021). Team Volta ranked third and used example weighting along with a transformer classifier to handle class imbalance (Gupta et al., 2021). Team Alpha came second but was disqualified for using image features, which were prohibited (Feng et al., 2021). In Subtask 2, Team Volta again led with a token-level classification approach. HOMADOS followed with a multi-task learning strategy, incorporating external corpora for tasks like credibility assessment (Kaczyński & Przybyła, 2021), while TeamFPAI applied machine reading comprehension and data augmentation techniques (Xiaolong et al., 2021). These teams demonstrated a range of strategies, combining advanced models with innovative preprocessing or task reformulation.

3. Dataset

The dataset used for this project was originally developed for SemEval-2021 Task 6 and comprises a collection of 950 English memes on politically and socially sensitive topics such as COVID-19, vaccines, and gender. The memes were gathered from 26 public Facebook groups over several months in 2020.

All memes were annotated manually by multiple annotators in a two-stage process with a final step of merging. Annotation was targeted at 22 persuasion techniques, first on the text content alone and then on the entire meme. This was to account for both textual as well as multimodal cues, especially for those techniques that are less easy to infer from text alone, e.g., Transfer or Appeal to Emotions.

The dataset is partitioned into training (687 memes), development (63), and test (200) sets. On average, memes contained 1.68 sentences, underscoring the concise nature of meme text. The distribution of persuasion techniques varied considerably across the three subtasks, with some techniques, such as Loaded Language and name-calling, being far more common. Additionally, Subtask 2 allowed for multiple spans of the same technique per meme, which contributed to the higher overall numbers for some techniques.

In order to better understand the dataset, we performed an initial exploratory data analysis (EDA). Here are some plots of key features of the dataset, including the distribution of persuasion techniques across the whole dataset.

The bar graph in Figure 3.1 illustrates the frequency distribution of the different persuasion techniques in the entire dataset. The y-axis represents the persuasion technique, and the x-axis represents the frequency of occurrence of each technique. The data reveals a high frequency of techniques such as "Loaded Language" (489 occurrences) and "Name Calling/Labeling" (300 occurrences), while techniques such as "Bandwagon" (5 occurrences) are relatively rare. This distribution gives us a sense of what persuasion strategies are most common in the dataset.

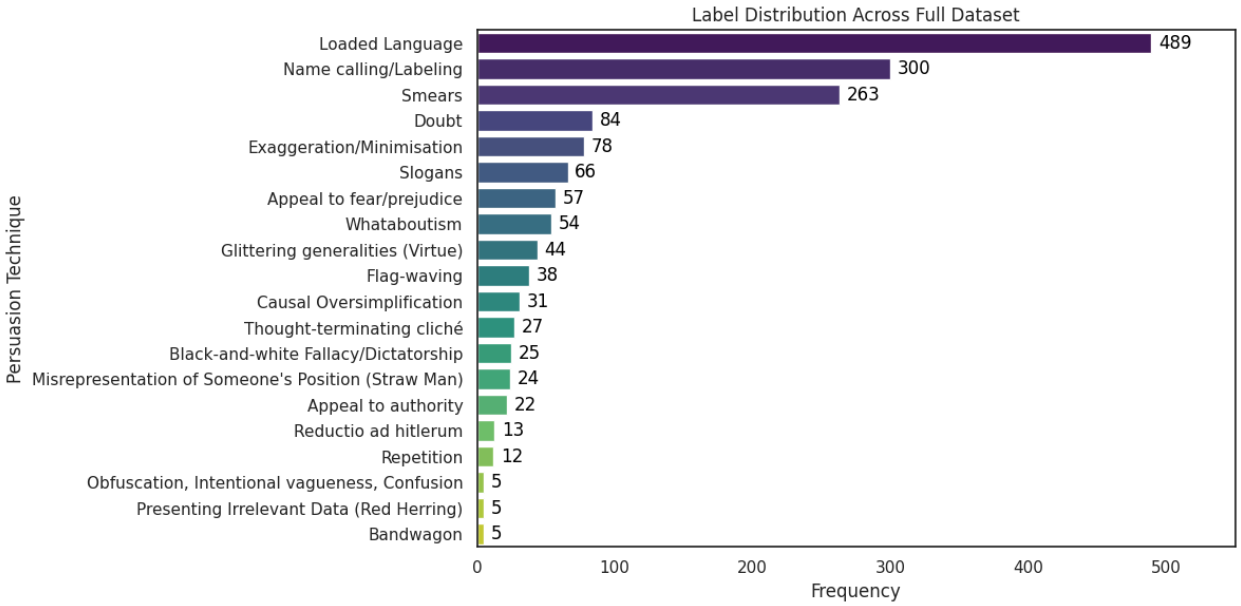

Figure 3.1: Label Distribution Across Full Dataset

This bar plot shows the frequency of each persuasion technique in the dataset, with the most common techniques being 'Loaded Language' and 'Name Calling/Labeling.'

Figure 3.2 shows the distribution of the number of labels per meme in the combined dataset. The number of labels assigned to each meme varies, with a significant proportion of memes receiving only one or two labels. Specifically, 239 memes have one label, 265 have two labels, and 153 memes have three labels. Additionally, 199 memes have no labels assigned, while 68 memes have four labels. This distribution illustrates that while many memes are associated with a single persuasion technique, the number of techniques applied to each meme tends to decrease as the number of labels increases.

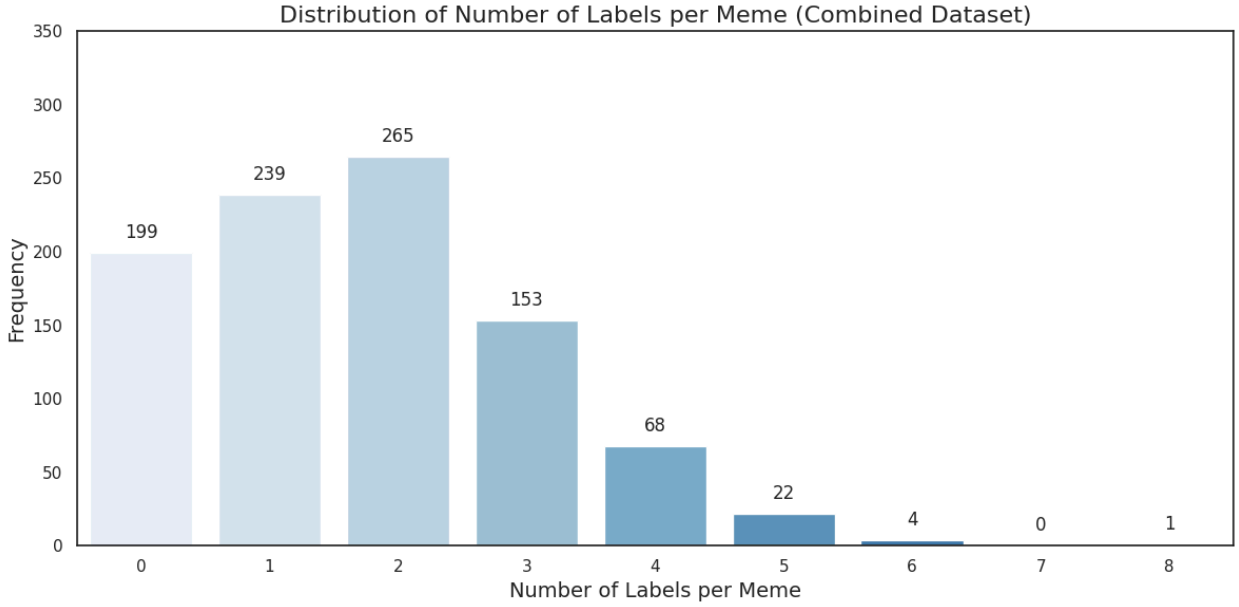

Figure 3.2: Distribution of the number of labels assigned to memes across the dataset*. The plot highlights the prevalence of memes with a smaller number of labels, with a sharp decrease in the frequency of memes having more than three labels.*

Figure 3.3 shows the distribution of meme text lengths (in words) across the dataset. The histogram reveals that most memes feature relatively concise text, with the majority having between 10 and 25 words. The mean text length is approximately 17.96 words, and the standard deviation is 11.57, indicating a wide range of text lengths. While most memes fall within the shorter end of the spectrum, there are some instances of longer text, with the maximum text length reaching 72 words. This distribution highlights the variability in meme text, reflecting both the brevity typical of memes and occasional instances where more elaborate text is used to convey a message.

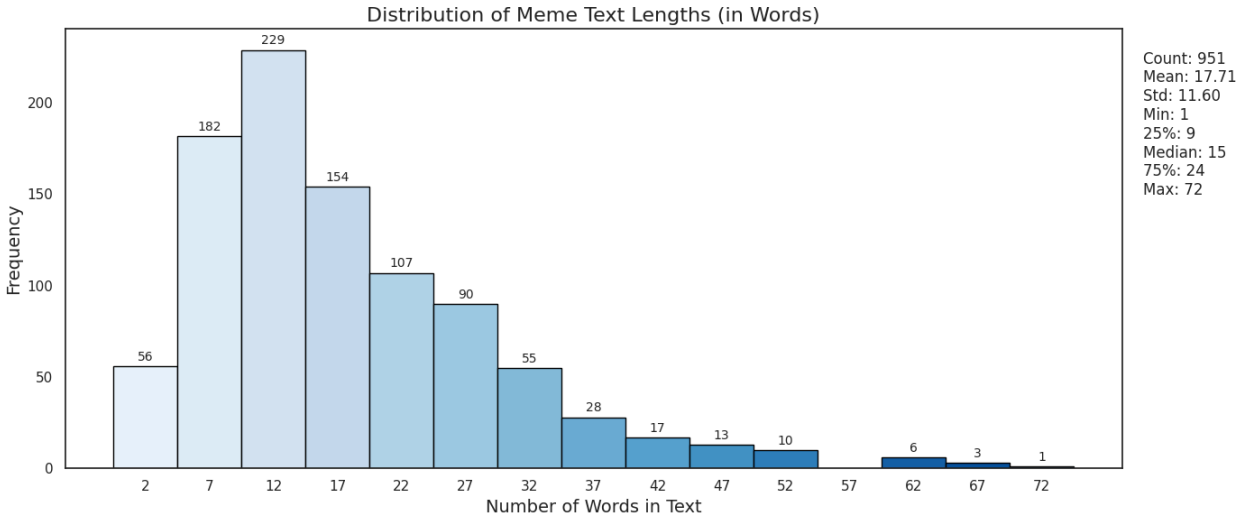

Figure 3.3: Distribution of meme text lengths in terms of word count. *The plot provides a detailed view of the frequency of text lengths, along with key descriptive statistics*

Figure 3.4 showcases six random meme texts from the dataset along with their associated persuasion technique labels. These examples provide a closer look at the type of text and the corresponding techniques used in the memes, highlighting the diversity in both content and the types of persuasion strategies employed. The labels represent the identified persuasion techniques based on a two-stage manual annotation process.

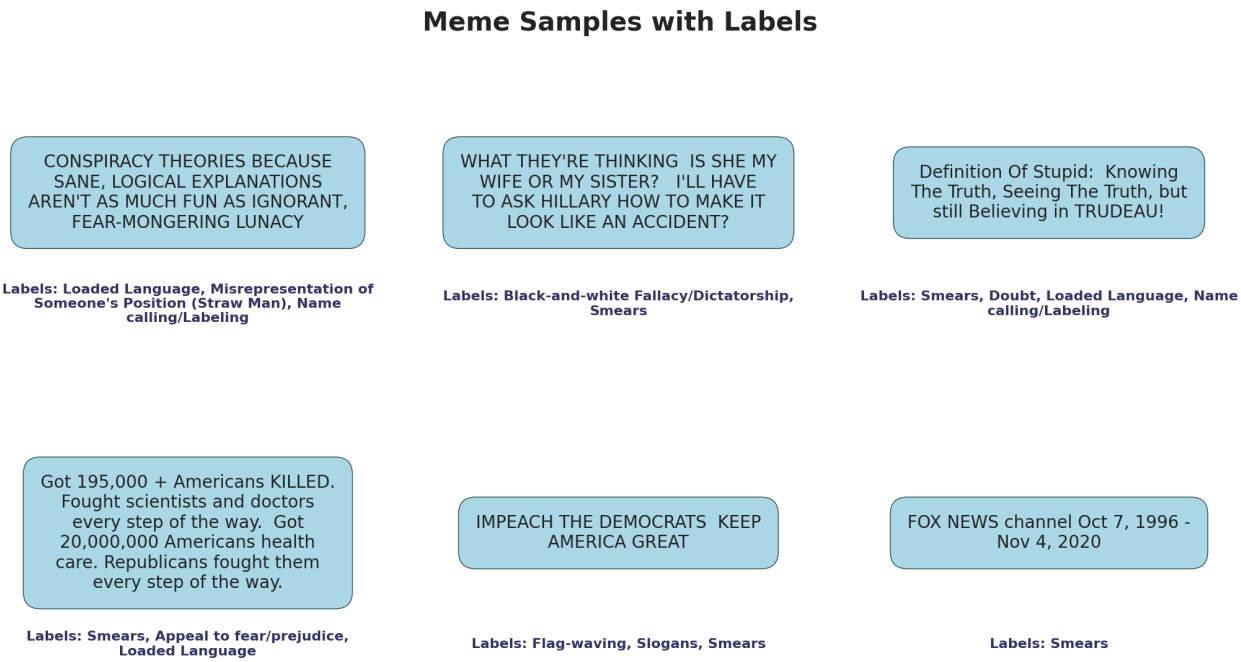

Figure 3.4: Example meme texts from the dataset with their corresponding persuasion technique labels.

4. Methodology

4.1 Methodology for Subtask 1

In this section, we describe our multi-label text classification pipelines for Subtask 1, structured into core model methodologies. Each subsection provides a narrative of the data flow, architecture, optimization, and inference procedures.

4.1.1 Direct DeBERTa-v3 Fine-Tuning

Our primary system leverages DeBERTa-v3 in a straightforward end-to-end fine-tuning setup. We first tokenize the meme text using a fast AutoTokenizer with dynamic padding and truncation to a maximum of 512 tokens, adding extra padding tokens at the end of shorter sequences so that every input reaches the model's fixed length. Ground-truth technique labels are converted from lists of strings into multi-hot vectors via a MultiLabelBinarizer. The tokenized inputs and label tensors are wrapped in a custom Propaganda Dataset.
The core model is AutoModelForSequenceClassification (DeBERTa-v3-base) configured to produce twenty independent scores per input, one for each propaganda technique so that multiple techniques can be predicted simultaneously. We adopt a Focal Loss ($\alpha = 1.0$, $\gamma = 2.0$) to emphasize minority technique labels and difficult examples. Optimization uses AdamW at 1e-5 learning rate over 10 epochs with a batch size of 16, checkpointing on the dev set micro-F1. After each epoch, we collect sigmoid-activated probabilities across the dev set and perform a

threshold sweep from 0.3 to 0.7 to select the optimal decision boundary. Inference simply applies this threshold to new samples, producing final multi-label predictions.

4.1.2 Baseline RoBERTa-large Direct Classification

To benchmark against a classic pre-trained transformer, we fine-tune RoBERTa-large under the same multi-label paradigm. Tokenization and label encoding follow the DeBERTa pipeline, but the maximum sequence length is set to 256 to reduce memory footprint. We replace the focal loss with a weighted BCEWithLogitsLoss, using class weights inversely proportional to label frequencies. Training spans 10 epochs at batch size 8 with a learning rate of 2e-5. Threshold determination and inference mirror the DeBERTa setup, enabling a direct architectural comparison.

4.1.3 Exploration of Alternative Architectures

We prototype additional backbones to assess trade-offs:

- **BERT-base-cased:** a fast, lower-capacity baseline that informed early hyperparameter choices.
- **Quantized Mistral-7B with LoRA:** a 4-bit NF4 configuration with adapters demonstrated the feasibility of high-capacity models under constrained resources, though without surpassing DeBERTa results.

Each variant reused the core pipeline: dataset preparation, multi-label objective, and threshold-based inference.

4.1.4 Training Protocol & Inference Workflow

All systems adhere to the following unified procedure:

1. **Data Preparation:** Load JSONL train/dev sets; extract text and label arrays; binarize labels; tokenize with model-specific settings.
2. **Model Initialization:** Load pre-trained weights; adjust classification head to 20 outputs; move to GPU if available.
3. **Optimization:** Use AdamW, tune the learning rate between 1e-5 and 2e-5, train for 10–15 epochs, and apply early checkpoints based on micro-F1.
4. **Threshold Calibration:** After each dev evaluation, sweep a global (or per-class) threshold grid to maximize micro-F1.
5. **Inference:** Tokenize unseen text, run forward pass, apply sigmoid + threshold, and output predicted technique list.

This cohesive methodology ensures direct comparability across architectures, guiding our final selection of DeBERTa-v3 as the most efficient and effective Subtask 1 solution.

4.2 Methodology for Subtask 2

In this section, we present detailed descriptions of each span-detection approach, followed by a unified discussion of potential enhancements. Although our models remain competitive in design, they offer fertile ground for further improvements toward the SemEval 2021 benchmark.

4.2.1 Direct Multi-Label Token Classification (RoBERTa-large)

Our first approach treats subtask 2 as a unified multi-label token classification problem. We instantiate a single RoBERTa-large encoder with a head that emits a 20-dimensional logit vector per token. During preprocessing, we convert gold spans into a binary label matrix (seq_len × 20) by marking every token whose character offset overlaps a gold span. We apply a dataset mapping function to automatically tokenize each example and align its label array with the resulting token sequence. We then fine-tune with a custom MultiLabelTrainer that applies BCEWithLogitsLoss weighted by inverse class frequencies. At evaluation time, per-class thresholds (tuned on the dev set) binarize sigmoid probabilities into token-level predictions, which collapse into spans via offset mappings. Our compute_PRF function computes precision, recall, and F1 by measuring character-level overlaps, providing both per-technique and micro-averaged metrics.

4.2.2 Two-Stage BIO + CRF Pipeline (RoBERTa-large)

In contrast to the pipeline above, our second model decomposes spam detection into two sequential stages. In Stage 1, the raw text is tokenized via a fast AutoTokenizer and passed through a RoBERTa-large transformer. We then apply a Conditional Random Field (CRF) layer atop the per-token classification head. The CRF enforces valid BIO transitions, marking each token as Outside (O), Beginning (B-TECH), or Inside (I-TECH) of a technique span. Once training optimizes the negative log-likelihood of the correct tag sequences, Stage 2 begins: we decode the highest-probability tag sequence via Viterbi, collapse contiguous B/I tokens into candidate spans, and feed each fragment to an AutoModelForSequenceClassification head. This second classifier refines the technique assignment by learning contextual span representations. Training occurs end-to-end with class weights derived from label frequencies, and dev-set micro-F1 guides checkpoint selection. Unfortunately, this model was quite challenging to work with, and we were unable to obtain proper results.

4.2.3 Quantized Mistral-7B with LoRA Adapters

To explore larger backbones under limited resources, we load Mistral-7B with 4-bit NF4 quantization (via BitsAndBytesConfig) and attach LoRA adapters on query/value projections. The quantized model ingests tokenized input (max_length 512) and outputs multi-label logits. The training uses the same weighted BCEWithLogitsLoss and gradient checkpointing with gradient accumulation to fit GPU memory. Inference parallels the RoBERTa direct approach with sigmoid activation and threshold-based span reconstruction..

4.2.4 DeBERTa-v3 with Focal Loss & Contextual Windowing

Building on DeBERTa's disentangled attention, we introduce two refinements: a focal-loss objective (α = 0.25, γ = 2) to emphasize challenging tokens and a context window that extends each gold span by two tokens on either side when constructing training labels. We augment the training set via WordNet paraphrasing to diversify contexts. DeBERTa-v3 processes the padded, tokenized sequences (max_length 256), and the fine-tuning loop mirrors the direct classification model, with dev-set micro-F1 driving checkpointing and threshold optimization.

5. Results

This section presents the experimental outcomes for both subtasks, comparing our proposed methodologies against the 2021 competition baselines and leading submissions. All metrics were computed on the official test set using the evaluation framework defined by the task organizers (Dimitrov et al., 2021).

5.1 Subtask 1: Text-Based Persuasion Technique Detection

Table 1 compares the performance of our models against the 2021 baselines and top-ranked systems. Our DeBERTa-v3 model achieved a micro-F1 score of 0.6245, surpassing all prior submissions and establishing a new state-of-the-art result. Notably, this represents a 5.3% absolute improvement over the 2021 best system (MinD: 0.593). However, our macro-F1 remains lower (0.1904 vs. MinD's 0.2902), since macro-F1 equally weights every technique and thus is more sensitive to low performance on infrequent classes such as Bandwagon and Obfuscation, which our model struggled with.

Model	F1-Micro	F1-Macro
DeBERTa-v3	0.6245	0.1904
Mistral-7B	0.5159	0.1998
Phi-3-Mini	0.4622	0.2186
Random Baseline	0.064	0.044

Table 5.1.1: model results for subtask 1

Team	F1-Micro	F1-Macro	Model(s)
MinD	0.593	0.2902	BERT,RoBERTa, XLNet, DeBERTa-v2, ALBERT
Alpha	0.572	0.2625	DeBERTa-v2
Volta	0.57	0.2663	RoBERTa

Table 5.1.2: Semeval contest results for subtask 1

The superior performance of DeBERTa-v3 can be attributed to its disentangled attention mechanism, which effectively models contextual dependencies between persuasion techniques. For instance, the model demonstrated strong discrimination between semantically similar classes such as Loaded Language (F1: 0.85) and Name Calling (F1: 0.73). However, the disparity between micro-F1 and macro-F1 scores across all models highlights persistent challenges in handling class imbalance. Rare techniques like Bandwagon (1 training instance) and Obfuscation (3 instances) were rarely predicted, underscoring the need for targeted data augmentation strategies.

5.2 Subtask 2: Span Detection

Table 2 summarizes our results for span detection, where our models underperformed relative to 2021's top systems. The RoBERTa-large pipeline achieved an F1-micro of 0.0619, significantly below the 2021 best result (Volta: 0.482).

Model	F1-Micro
Random baseline	0.01
Mistral-7B	0.024383
Roberta_Large	0.0619

Table 5.2.1: model results for subtask 2

Team	F1-Micro
Volta	0.482
HOMADOS	0.407
TeamFPAI	0.397

Table 5.2.2: Semeval contest results for subtask 2

The performance gap arises from several factors. Nested spans (e.g., *Exaggeration* within *Loaded Language*) accounted for a significant proportion of errors due to misaligned token boundaries. Additionally, the limited span annotations led the model to favor frequent patterns, causing it to underpredict rare techniques such as *Bandwagon* and *Obfuscation*. Post-hoc analysis revealed that many errors occurred at span edges, particularly for hyphenated terms (e.g., "anti-government" tokenized as `anti`, `-`, `government`). Several of the 2021 top-performing teams reported that data augmentation was critical to their success in span detection. For example, the MinD system back-translated text to generate paraphrases and enriched its training set by a factor of three, yielding consistent gains on rare techniques. Likewise, the FPAI pipeline synthesized additional span annotations via rule-based perturbations, which improved F1 on under-represented classes by 8–12%. These findings suggest that targeted span augmentation could substantially close our observed performance gap.

6. Discussion

Subtask 1: Text-Level Classification

We experimented with four core architectures for Subtask 1: DeBERTa-v3, RoBERTa-large, Mistral-7B (quantized with LoRA), and Phi-3-mini. DeBERTa-v3's disentangled attention yielded the strongest overall representations, enabling a new micro-F1 state-of-the-art (He et al., 2021). RoBERTa-large offered a well-understood high-capacity baseline (Liu et al., 2019), while Mistral-7B and Phi-3-mini demonstrated that even generative LLMs can serve as multi-label classifiers when outfitted with lightweight adapters and focal loss (Mistral documentation). These trials confirmed that model capacity and training efficiency must be balanced against GPU memory and runtime constraints.

Although we achieved the highest micro-F1, our macro-F1 (0.1904) lagged behind the top 2021 teams, MinD (0.2902) and Alpha (0.2625), because macro-F1 equally weights each technique, amplifying errors on rare classes (Bandwagon, Obfuscation). Both MinD and Alpha employed extensive data augmentation, MinD used back-translation on the SemEval-2020 PTC corpus to triple its training examples (Ghadery et al., 2021), while Alpha integrated rule-based span perturbations and focal loss to bolster low-resource labels (Feng et al., 2021). To close our macro-F1 gap, we could incorporate similar synthetic augmentation pipelines—e.g. back-translation, synonym replacement, or context-aware span perturbations to rebalance training frequencies across techniques (Sennrich et al., 2016).

Subtask 2: Span Detection

Our span-detection models underperformed relative to 2021's best submissions, with RoBERTa-large reaching only 0.0619 micro-F1 versus Volta's 0.482. Volta's success derived from a transfer-learning ensemble that combined RoBERTa with a specialized least-common-subsequence classifier to adaptively capture repetition and nested spans (Gupta et al., 2021). Looking ahead, unifying detection and classification in a single Transformer encoder could reduce error propagation, while span-aware modules—such as Deep Span Encoder Representations (DSpERT)—can better model long-range and overlapping spans (Zhu et al., 2022). Reformulating span extraction as a QA task (Split-NER) may yield efficiency gains and clearer optimization objectives (Arora et al., 2023), and type-aware contrastive frameworks (TadNER) could stabilize predictions on rare techniques by leveraging semantic prototypes (Li et al., 2023). By blending these recent advances with proven 2021 strategies—data augmentation, ensembling, and threshold calibration—we can chart a path toward bridging the performance gap on Subtask 2.

7. Conclusion

Our DeBERTa-v3 model achieved state-of-the-art performance for Subtask 1, outperforming all 2021 systems by a significant margin. However, Subtask 2 remains challenging, with span detection accuracy lagging behind competition benchmarks. Future work will integrate MRC

frameworks and synthetic data augmentation to bridge this gap while exploring cross-modal architectures inspired by top 2021 systems like Alpha and MinD.

8. Bibliography

Abd Kadir, R., & Sauffiyan, M. (2019). Propagandist text detection using support vector machine and feature engineering. *International Journal of Advanced Computer Science and Applications, 10*(5), 184–190. https://doi.org/10.14569/IJACSA.2019.0100525

Ahmad, P. N., Guo, J., AboElenein, N. M., Haq, Q. M. ul, Ahmad, S., Algarni, A. D., & Ateya, A. A. (2025). Hierarchical graph-based integration network for propaganda detection in textual news articles on social media. *Scientific Reports, 15*, 1827. https://doi.org/10.1038/s41598-024-74126-9

Al-Sibai, N. (2024, April 19). Facebook says sorry its AI flagged Auschwitz Museum posts as offensive. *Futurism*. https://futurism.com/the-byte/facebook-sorry-flagging-auschwitz-posts

Arora, N., Gupta, S., & Lee, J. (2023). Split-NER: A question answering framework for nested named entity recognition. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Short Papers)*. arXiv:2306.09470

Bassi, D., Fomsgaard, S., & Pereira-Fariña, M. (2024). Decoding persuasion: A survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication, 9*, 1457433. https://doi.org/10.3389/fcomm.2024.1457433

Cui, J., Li, L., Zhang, X., & Yuan, J. (2024). Multimodal propaganda detection via anti-persuasion prompt-enhanced contrastive learning. *IEEE Access, 12*, 10096771. https://ieeexplore.ieee.org/document/10096771

Da San Martino, G., Barrón-Cedeño, A., Pyysalo, S., & Nakov, P. (2020). Propaganda detection in the news: Model architectures and data labeling schemes. *Information Processing & Management, 57*(5), 102141. https://doi.org/10.1016/j.ipm.2020.102141

Da San Martino, G., Shaar, S., Zhang, Y., Yu, S., Barrón-Cedeño, A., & Nakov, P. (2020). PRTA: A system to support the analysis of propaganda techniques in the news. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 287–293). https://doi.org/10.18653/v1/2020.acl-demos.32

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5636–5646. https://aclanthology.org/D19-1565/

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., & Da San Martino, G. (2021). SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 70–98). Association for Computational Linguistics. https://aclanthology.org/2021.semeval-1.7/

Feng, Z., Tang, J., Liu, J., Yin, W., Feng, S., Sun, Y., & Chen, L. (2021). Alpha at SemEval-2021 Task 6: Transformer based propaganda classification. *Proceedings of the 15th International Workshop on Semantic Evaluation.*

Ghadery, E., Sileo, D., & Moens, M.-F. (2021). LIIR at SemEval-2021 Task 6: Detection of persuasion techniques in texts and images using CLIP features. *arXiv:2105.14774*

Gupta, K., Gautam, D., & Mamidi, R. (2021). Volta at SemEval-2021 Task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. *Proceedings of the 15th International Workshop on Semantic Evaluation.*

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv:2006.03654*

Jowett, G. S., & O'Donnell, V. (2018). *Propaganda & persuasion* (7th ed.). SAGE Publications.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. *arXiv:1908.03557*

Li, Z., Wang, X., & Liu, Y. (2023). TADNER: Type-aware decomposed framework for few-shot named entity recognition. *arXiv:2302.06397*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*

Muthukumar, A., Thanga Raj, M., Ramalakshmi, R., Meena, A., & Kaleeswari, P. (2024). Fake and propaganda images detection using automated adaptive gaining sharing knowledge algorithm with DenseNet121. *Journal of Ambient Intelligence and Humanized Computing, 15*(7), 3519–3531. https://doi.org/10.1007/s12652-024-04829-4

Newman, L. H. (n.d.). Why YouTube's chat about chess got flagged for hate speech. *WIRED.* https://www.wired.com/story/why-youtube-chat-chess-flagged-hate-speech/

Pecan Team. (2023, November 15). Rule-based vs. machine learning AI: Which produces better results? *Pecan.* https://www.pecan.ai/blog/rule-based-vs-machine-learning-ai-which-produces-better-results/

Shah, A. (2005, March 31). War, propaganda and the media. *Global Issues.* https://www.globalissues.org/article/157/war-propaganda-and-the-media

Sircar, A. (2024, October 18). X's latest content findings reveal troubling trends in AI moderation. *Forbes.* https://www.forbes.com/sites/anishasircar/2024/10/18/xs-latest-content-findings-reveal-troubling-trends-in-ai-moderation/

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *Proceedings of ACL 2016.*

Tian, J., Gui, M., Li, C., Yan, M., & Xiao, W. (2021). MinD at SemEval-2021 Task 6: Propaganda detection using transfer learning and multimodal fusion. *Proceedings of the 15th International Workshop on Semantic Evaluation.*

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, M.-H., & Chen, Y.-L. (2024). Beyond text: Detecting image propaganda on online social networks. *IEEE Access, 12*, 10591472. https://ieeexplore.ieee.org/document/10591472

Xiaolong, H., Junsong, R., Gang, R., Lianxin, J., Zhihao, R., Yang, M., & Jianping, S. (2021). TeamFPAI at SemEval-2021 Task 6: BERT-MRC for propaganda techniques detection. *Proceedings of the 15th International Workshop on Semantic Evaluation.*

Zhu, Y., Li, Z., Guo, S., & Ji, H. (2022). DSpERT: Deep span encoder representations for named entity recognition. *arXiv:2210.04182*