Multi-Approach Propaganda Detection: Technique Classification and Span Identification in Text Snippets

Shashank Sathyan

1 Introduction

Propaganda, the systematic dissemination of information to influence opinions, attitudes, or actions, has been employed throughout human history. From ancient war manifestos to World War-era posters, propaganda has served as a powerful tool for shaping public opinion and behavior. In today's digital landscape, propaganda has evolved far beyond printed materials and radio broadcasts. Internet and social media have dramatically transformed how information spreads, creating unprecedented opportunities for the rapid and widespread dissemination of propaganda. Modern propaganda can be subtle, targeted, and difficult to distinguish from legitimate information. Propaganda can be classified into multiple techniques. (Martino et al. 2020) classifies propaganda into 14 major techniques some of which are combined. In this paper we look at 8 techniques and one more that is 'not propaganda' to identify sentences that are not propaganda. The techniques looked at are: (1) flag waving (2) appeal to fear prejudice (3) causal simplification (4) doubt (5) exaggeration,minimisation (6) loaded language (7) name calling,labeling and (8) repetition. The paper addresses two fundamental tasks in propaganda detection: (1) classifying the specific propaganda technique used within an identified propaganda span, and (2) detecting both the span and technique of propaganda within a larger text. For the first task, we implement and evaluate three distinct approaches: (1) fine-tuning RoBERTa-large with extracted propaganda spans, (2) reformulating classification as a text generation task using T5, and (3) a traditional machine learning approach using Support Vector Machines with TF-IDF features. For the second, more complex task of identifying propaganda spans within sentences, we explore two approaches: (1) a sequence labeling method using RoBERTa with BIO tagging, and (2) span classification approach using RoBERTa with negative sampling. These diverse methodologies provide complementary strengths in addressing the challenges of propaganda detection.

2 Methodology

2.1 Task 1: Technique Classfication

For this task, given the sentence with the propaganda span we predict the type of propaganda technique being used.

2.1.1 ROBERTa

Transformer-based language models have revolutionized natural language processing (NLP) in recent years. Among these models, RoBERTa (Robustly Optimized BERT Pre-training Approach) (Liu et al. 2019) represents a significant improvement over BERT (Bidirectional Encoder Representations from Transformers). Developed by researchers at Facebook AI in 2019, RoBERTa builds upon BERT's architecture while implementing key optimization improvements including training on larger datasets, removing the next sentence prediction objective, using larger batch sizes, and employing a dynamic masking pattern. For our propaganda technique classification task, we followed an approach similar to (Abdullah et al. 2022). Their research demonstrated the effectiveness of transformer-based models, particularly RoBERTa, in identifying subtle propaganda techniques in news text. Building on their methodology, we implemented a fine-tuning approach using the RoBERTa-large pre-trained model. We extracted propaganda spans from the provided dataset, where each span was annotated with one of nine propaganda techniques. Each span was marked with <BOS> and <EOS> tags within its sentential context. Following (Abdullah et al. 2022) we applied several preprocessing steps to normalize the text:

1. Expanding contractions (e.g., ”don't” → ”do not”)
2. Replacing abbreviations with their full forms (e.g., ”U.S.” → ”United States of America”)
3. Converting text to lowercase
4. Removing redundant whitespace

This preprocessing strategy helps standardize the input and reduce noise that might otherwise impact the model's performance. We implemented RoBERTa-large as our classification model, which consists of 24 transformer layers, 16 attention heads, and 355M parameters. Following the approach in (Abdullah et al. 2022) we utilized the RobertaForSequenceClassification implementation, which adds a classification head on top of the transformer architecture for sequence-level classification tasks. Input spans were tokenized using the RobertaTokenizer with special tokens ([CLS] and [SEP]) added as markers. All sequences were truncated or padded to a maximum length of 120 tokens, with appropriate attention masks applied to ignore padding tokens during processing. The final hidden state of the [CLS] token serves as the aggregate sequence representation for classification purposes.

The model was trained with almost all parameters kept the same from the paper (Abdullah et al. 2022) except varying the learning rate with values 1e-5 and 2e-5.

2.1.2 T5 Model

For the second approach, we take inspiration from the appraoch taken in (Sprenkamp et al. 2023) where propaganda techniques were classified using both GPT-3 and GPT-4 models. In our approach, we look at fine-tuning the T5-base model (Raffel et al. 2020) for the task. This approach reformulates the classification task as a text generation problem. We preprocessed the propaganda dataset by transforming each instance into a structured prompt. The propaganda spans were extracted from their contexts and highlighted using special markers ([SPAN] and [/SPAN]). Each input was formatted as:

Hyper-parameter	Value
Number of epochs	4
Batch size	4
Learning rate	1e-5, 2e-5
Patience	2
Epsiolon values	(1e-30, 1e-3)
Clip threshold	1.0
Decay rate	-0.8
warmup_init	False
Optimizer	Adafactor
Loss function	CrossEntropyLoss

Table 1: Fine tuned parmaters for RoBERTa model

Article: [context with highlighted span]

Prompt :Identify the propaganda technique used in the text between `[SPAN]` **and** `[/SPAN]` **markers.**

Options: doubt, repetition, appeal_to_fear_prejudice, ...

Answer: [selected technique]

To address the class imbalance present in the dataset, we implemented a weighted random sampling approach where each class had an equal probability of being selected during training, regardless of its frequency in the dataset. We also define the hyper parameter setting for our model in Table 2.

Hyperparameter	Value
Learning rate	3e-5
Optimizer	AdamW
Batch size	8
Epochs	3
Max input sequence length	150 tokens
Max output sequence length	25 tokens

Table 2: Fine tuned parameters for T5 model

2.1.3 Support Vector Machine

For our third approach, we implemented a Support Vector Machine (SVM) classifier with Term Frequency-Inverse Document Frequency (TF-IDF) feature representation to classify propaganda techniques in text spans. SVMs are particularly effective for text classification tasks as they can handle high-dimensional feature spaces efficiently and perform well on imbalanced datasets (Cortes and Vapnik 1995). The training data consisted of labeled text spans with propaganda techniques and their surrounding context. We extracted the text between the ¡BOS¿ and ¡EOS¿ tags to focus on the specific span labeled with a propaganda technique. This approach ensures the model learns to classify based on the exact text that contains the propagandistic content rather than surrounding context. We employed TF-IDF vectorization to convert the text spans into numerical features. This approach captures both the importance of terms within a document and their discriminative power across the corpus. The TF-IDF configuration is as follows:

1. N-gram range of (1, 2) to capture both individual words and meaningful pairs
2. Maximum of 10,000 features to balance model complexity and computational efficiency
3. Minimum document frequency of 3 to filter out extremely rare terms
4. Sublinear TF scaling to dampen the effect of term frequency

We implemented a pipeline consisting of TF-IDF vectorization followed by a Linear SVM classifier. To address class imbalance, we employed balanced class weights that automatically adjust weights inversely proportional to class frequencies. For parameter optimization, we performed a grid search with 5-fold cross-validation, exploring:

1. Different n-gram ranges: (1,1), (1,2), and (1,3)
2. Various feature set sizes: 5,000, 10,000, and 15,000 features
3. Multiple regularization strengths (C parameter): 0.01, 0.1, 1.0, and 10.0

2.2 Task 2: Span Identification and Technique Classification

For Task 2, we implement two different approaches to identify the span in a given snippet however we use Model B from Approach 1 of Task 1 i.e the RoBERTa model to further classify the technique in the span.

2.2.1 RoBERTA with BIO tagging

We implemented a sequence labeling approach to detect propaganda spans in sentences. Our methodology draws inspiration from (Chernyavskiy et al. 2020) which demonstrated the effectiveness of combining RoBERTa with CRF for span detection tasks. However, in our approach we don't implement the CRF layer rather just use the RoBERTa base model with BIO tagging. The training data consisted of sentences with propaganda spans marked by <BOS> and <EOS> tags. Our first step involved transforming this data into a format suitable for token-level classification. We implemented a custom PropagandaDataset class that processes each text sample by:

1. Locating the <BOS> and <EOS> markers to identify propaganda spans
2. Removing these markers while preserving their position information
3. Tokenizing the cleaned text using the RoBERTa tokenizer
4. Converting character-level span annotations to token-level BIO (Beginning-Inside-Outside) tags

This conversion from character-level spans to token-level tags was particularly challenging due to tokenization boundaries not always aligning with word boundaries. Our implementation carefully handles these edge cases, ensuring that tokens are correctly labeled even when they partially overlap with propaganda spans. The BIO tagging scheme we employed consists of:

1. B-PROP: Assigned to the first token of a propaganda span
2. I-PROP: Assigned to continuation tokens within a propaganda span
3. O: Assigned to tokens outside any propaganda span

Our model architecture consists of :

1. RoBERTa-base encoder (125 million parameters)
2. A token classification head implemented as a linear layer that projects the hidden states to the label space (3 classes i.e The B-Prop, I-Prop and O)
3. Cross-entropy loss function for training

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	8
Training Epochs	3
Maximum Sequence Length	256 tokens
Random Seed	42
Train-Validation Split	80/20
Model Architecture	RoBERTa-base
Number of Classes	3 (BIO tagging)
Loss Function	Cross-entropy

Table 3: Model Hyper parameters

A critical component of our system is the conversion of token-level predictions back to character-level spans, which involves several steps:

1. Token-to-Span Conversion: Converting sequences of B-PROP and I-PROP tokens back to continuous character spans using the offset mapping provided by the tokenizer.
2. Span Refinement Pipeline:
   (a) Merging Overlapping/Adjacent Spans: Combining spans that overlap or are separated by at most 2 characters, addressing fragmentation issues caused by tokenization.
   (b) Boundary Refinement: Adjusting span boundaries to ensure they start and end with alphanumeric characters, removing trailing punctuation or whitespace.
   (c) Quotation Mark Handling: Expanding spans to include quotation marks at boundaries, preserving important semantic markers.

(d) Length Filtering: Removing spans shorter than 2 characters to reduce false positives.

3. Span Matching Criteria: During evaluation, we consider a predicted span correct if it overlaps with a true span by at least 50% of the length of the shorter span, providing a balance between strict and lenient evaluation.

We conducted a comprehensive ablation study to assess the impact of each post-processing component, comparing different configurations to determine the optimal approach for span extraction.

2.2.2 RoBERTa with Negative Sampling

For this approach we implemented a sequence pair classification approach that treats span detection as a binary classification problem. This approach is inspired by work in named entity recognition and question answering that frames span detection as classification over candidate spans. For each positive example in the training set, we did the following:

1. Extracted the full sentence context by removing the BOS/EOS markers.
2. Identified the gold span marked between BOS/EOS tokens.
3. Generated negative examples by sampling other spans from the same sentence.
4. Used a negative-to-positive ratio of 5:1 to address class imbalance.
5. Limited candidate spans to a maximum of 10 words to control computational complexity.

This approach allowed the model to learn what constitutes propaganda by seeing both positive examples (actual propaganda spans) and negative examples (spans that are not propaganda) within the same contexts. We implemented a sequence pair classification model based on RoBERTa (Liu et al. 2019), a robustly optimized BERT-based model that has shown strong performance on various NLP tasks. The model takes two inputs:

1. The complete sentence context
2. The candidate span

The model architecture consists of :

1. A pre-trained RoBERTa-base encoder (125M parameters)
2. A classification head to predict the propaganda technique label (including ”not_propaganda”)
3. Cross-entropy loss function for multi-class classification

This approach allows the model to learn contextual relationships between spans and their surrounding text. During inference, we approached span detection as a span extraction problem:

1. Generated all possible candidate spans up to 10 words long from each test sentence
2. Processed these spans in batches of 64 to manage memory constraints

Parameter	**Value**
Learning rate	2e-5
Weight decay	0.01
Batch size (training)	4
Batch size (evaluation)	8
Training epochs	3
Maximum sequence length	128 tokens
Optimizer	AdamW with linear learning rate scheduler

Table 4: Model Hyperparameters

3. For each candidate, paired it with the full sentence context and fed it to the model

4. Applied a softmax over the class probabilities and masked out the ”not_propaganda” class

5. Selected the span with the highest propaganda class probability as the predicted span

3 Results and Error Analysis

3.1 Task 1: Technique Classification

Below are the results and the error analysis of each approach implemented in completing Task 1.

3.1.1 RoBERTa

We evaluated two variants of our RoBERTa-large model, differing only in the learning rate used during fine-tuning: 1e-5 (Model A) and 2e-5 (Model B).

	Model A (1e-5)	Model B (2e-5)
Micro F1	0.7500	0.7812
Macro F1	0.6368	0.6694
Accuracy	0.7500	0.7812

Table 5: Overall Performance Comparison

The higher learning rate (Model B) yielded superior performance across all metrics, with a 3.12 percentage point improvement in micro F1 and a 3.26 percentage point improvement in macro F1. This suggests that the larger learning rate allowed the model to better adapt to the propaganda classification task without overfitting.

Interestingly, while Model B performed better overall, performance varied by class. The most dramatic improvement occurred in the ”repetition” class, where Model B achieved a 25.54 percentage point increase in F1 score (0.2581 to 0.5135). Similarly, ”doubt” saw a substantial 13.97 percentage point improvement. However, Model B underperformed Model A on three classes: ”flag_waving” (-4.32 percentage points), ”exaggeration,minimisation” (-11.41 percentage points), and ”name_calling,labeling” (-6.78 percentage points). This suggests that the optimal learning rate

Propaganda Technique	**Model A (1e-5)**	**Model B (2e-5)**	**Difference**
flag_waving	0.8387	0.7955	-0.0432
appeal_to_fear_prejudice	0.6286	0.6316	+0.0030
causal_oversimplification	0.6061	0.7059	+0.0998
doubt	0.5185	0.6582	+0.1397
exaggeration,minimisation	0.7317	0.6176	-0.1141
loaded_language	0.5517	0.5556	+0.0039
name_calling,labeling	0.7027	0.6349	-0.0678
repetition	0.2581	0.5135	+0.2554
not_propaganda	0.8952	0.9116	+0.0164
Micro F1	0.7500	0.7812	+0.0312
Macro F1	0.6368	0.6694	+0.0326

Table 6: F1 scores comparison between models with different learning rates

may vary by class, potentially due to differences in the complexity of patterns that need to be recognized.

Our error analysis revealed several key factors affecting model performance. Class imbalance significantly impacted results, with both models performing best on the majority ”not_propaganda” class (F1 scores of 0.8952 and 0.9116 for Models A and B respectively), while performance on other classes generally correlated with their frequency in the training data. The ”repetition” technique proved particularly challenging for both models, though Model B (F1=0.5135) substantially outperformed Model A (F1=0.2581), suggesting this technique's detection benefits from higher learning rates due to its reliance on broader context beyond the model's 120 token limitation. We observed consistent confusion patterns across both models, particularly between ”loaded_language” and ”name_calling,labeling” due to their shared characteristics of emotionally charged language, and between ”appeal_to_fear_prejudice” and ”exaggeration,minimisation.” Finally, the varying impact of learning rate across propaganda techniques indicates differences in their learning complexity; techniques with clearer lexical signals (like ”flag_waving”) performed well even with lower learning rates, while more contextually dependent techniques (like ”repetition”) showed marked improvement with higher learning rates

3.1.2 T5 model

The T5-based approach for classification of the propaganda technique yielded poor performance metrics across all evaluation measures. Table 5 presents the detailed performance breakdown by propaganda technique. Overall, the model achieved a micro-average F1 score of 0.0724, a macro-average F1 score of 0.0151, and a weighted-average F1 score of 0.0101. The model's overall accuracy was 0.07, indicating that it correctly classified only 7% of the test instances. Analysis of per-class performance reveals that the model essentially collapsed to predicting primarily one class - ”appeal_to_fear_prejudice” - which achieved a recall of 0.98 but a precision of only 0.07, resulting in an F1 score of 0.14. For all other propaganda techniques, the model failed to make any correct predictions, yielding precision, recall, and F1 scores of 0.00.

Technique	Precision	Recall	F1-score	Support
appeal_to_fear_prejudice	0.07	0.98	0.14	43
causal_oversimplification	0.00	0.00	0.00	31
doubt	0.00	0.00	0.00	38
exaggeration,minimisation	0.00	0.00	0.00	28
flag_waving	0.00	0.00	0.00	39
loaded_language	0.00	0.00	0.00	37
name_calling,labeling	0.00	0.00	0.00	31
not_propaganda	0.00	0.00	0.00	301
repetition	0.00	0.00	0.00	32
accuracy			0.07	580
macro avg	0.01	0.11	0.02	580
weighted avg	0.01	0.07	0.01	580

Table 7: Classification Report for T5-based Propaganda Technique Classification

Metric	Score
Micro-average F1	0.0724
Macro-average F1	0.0151
Weighted-average F1	0.0101

Table 8: Overall Performance Metrics for T5-based Propaganda Classification

Our T5-based approach exhibited a form of classifier degeneration, similar to phenomena observed in prior work (Gururangan et al. 2018), where despite implementing class balancing techniques, the model converged to predominantly predicting a single class. This pattern of degeneration in fine-tuned transformer models often occurs when the model exploits statistical patterns in the training data rather than learning the underlying task (Devlin et al. 2019). The classification report reveals this stark pattern, with the model achieving 0.98 recall but only 0.07 precision for ”appeal_to_fear_prejudice” class, while completely failing (0.00 F1 scores) on all other propaganda categories. This collapse into essentially a single-class predictor indicates fundamental issues with our approach. Several interrelated factors likely contributed to this failure. First, reformulating propaganda classification as text generation may have inadvertently simplified the task, causing the model to memorize common outputs rather than learn discriminative features. Second, though we implemented weighted random sampling to address class imbalance, this approach proved ineffective as evidenced by the model's inability to identify any instances of minority classes. The limited training period of three epochs may have been insufficient for the model to capture the nuanced distinctions between propaganda techniques, particularly given the task's inherent complexity. Furthermore, our prompt design, which presented all possible techniques as options, might have created interference between class representations rather than providing helpful context. The hyperparameter configuration—including learning rate (3e-5), batch size (8), and sequence length limits—potentially compounded these issues, as T5 models are known to be sensitive to parameter selection when fine-tuned on specialized tasks with limited data (Raffel et al. 2020). The model's behavior of high recall but low precision for a single class suggests it recognized general propaganda patterns but lacked the capacity to distinguish between specific techniques. This phenomenon aligns with observations from (Gururangan et al. 2018), who noted that models can exploit dataset artifacts rather than learning the intended task.

3.1.3 Support Vector Machine

The SVM model with TF-IDF features demonstrated moderate effectiveness in classifying propaganda techniques, achieving a macro-averaged F1 score of 0.2966 and a micro-averaged F1 score of 0.5172 on the test set. Performance varied considerably across different propaganda techniques. The model performed relatively well on identifying ”flag_waving” (F1 = 0.4719) and ”not_propaganda” (F1 = 0.7247), but struggled with classes like ”loaded_language” (F1 = 0.0000) and ”name_calling,labeling” (F1 = 0.1404).

Propaganda Technique	Precision	Recall	F1-score	Support
appeal_to_fear_prejudice	0.3261	0.3488	0.3371	43
causal_oversimplification	0.3125	0.3226	0.3175	31
doubt	0.2564	0.2632	0.2597	38
exaggeration,minimisation	0.2174	0.1786	0.1961	28
flag_waving	0.4200	0.5385	0.4719	39
loaded_language	0.0000	0.0000	0.0000	37
name_calling,labeling	0.1538	0.1290	0.1404	31
not_propaganda	0.6918	0.7608	0.7247	301
repetition	0.2727	0.1875	0.2222	32

Table 9: Performance metrics for propaganda technique classification using SVM with TF-IDF

Our error analysis revealed that the SVM classifier exhibited notable performance disparities across the propaganda techniques, with implications for both the model architecture and feature representation choices. Analysis of the classification results reveals a significant impact of class imbalance, with the dominant ”not_propaganda” class (301 instances) achieving the highest F1 score (0.7247), while minority classes generally performed worse despite the implementation of balanced class weights. Most strikingly, the model completely failed to identify ”loaded_language” instances (F1 = 0.0000), suggesting a fundamental limitation of TF-IDF's lexical features in capturing the connotative and emotional aspects that characterize this propaganda technique. The poor performance on ”name_calling,labeling” (F1 = 0.1404) can be attributed to the diverse lexical manifestations of pejoratives across different contexts and domains, making it difficult for n-gram features to generalize effectively. Conversely, techniques like ”flag_waving” (F1 = 0.4719) and ”appeal_to_fear_prejudice” (F1 = 0.3371) achieved moderate performance, likely due to their more consistent phraseology and lexical patterns that TF-IDF can effectively capture. The model's inadequate performance on ”repetition” (F1 = 0.2222) highlights a critical limitation of span-level classification, as identifying repetition inherently requires document-level context beyond individual text spans. The substantial gap between micro-averaged (0.5172) and macro-averaged (0.2966) F1 scores further underscores the model's struggles with class imbalance. These findings suggest that while SVMs with TF-IDF features provide a reasonable baseline for propaganda technique classification, they fail to capture the semantic nuances and contextual dependencies that characterize certain propaganda techniques.

3.2 Task 2: Span Identification and Technique Classification

Below are the results and error analysis of each approach implemented for Task 2.

3.2.1 RoBERTa with BIO tagging

Our RoBERTa-based sequence labeling approach achieved exceptional performance on the propaganda span identification task. After the final epoch of training, the model reached impressive metrics with a precision of 0.9959, recall of 0.9938, and F1 score of 0.9948 Table 9.

The evaluation on the test set (Table 10) demonstrated consistent performance with precision, recall, and F1 scores all at 0.9914, indicating robust generalization capabilities. These results significantly outperform previously reported benchmarks for propaganda span detection tasks.

Metric	**Value**
Precision	0.9914
Recall	0.9914
F1 Score	0.9914

Table 10: Test Set Evaluation Results

Metric	**Value**
Precision	0.9959
Recall	0.9938
F1 Score	0.9948

Table 11: Model Performance After Final Training Epoch

To ensure the reliability of our approach, we conducted a 5-fold cross-validation evaluation.

Fold	**Precision**	**Recall**	**F1 Score**
1	0.9816	0.9959	0.9887
2	1.0000	0.9896	0.9948
3	1.0000	0.9959	0.9979
4	0.9979	0.9938	0.9959
5	0.9958	0.9938	0.9948
Average	0.9951	0.9938	0.9944

Table 12: 5-Fold Cross-Validation Results

The average metrics across all folds were precision 0.9951, recall 0.9938, and F1 score 0.9944 (Table 11), with a notably small standard deviation (±0.0034 for F1 score). This consistency across folds indicates that our model's performance is not dependent on a specific data split, supporting the robustness of our approach. To further evaluate our approach we perform an Ablation study the results of which are shown in Table 12.

Configuration	Precision	Recall	F1 Score	Change
Base model (with post-processing)	0.9979	0.9938	0.9959	-
No post-processing	0.9979	0.9959	0.9969	+0.10%
Strict span matching	0.8503	0.8468	0.8485	-14.79%
No post-processing + Strict matching	0.9896	0.9876	0.9886	-0.73%

Table 13: Comparison of model performance under different configurations

Interestingly, the model without post-processing achieved the best performance with an F1 score of 0.9969, showing a slight improvement over the baseline. This suggests that RoBERTa already excels at identifying token boundaries for propaganda spans, and the post-processing steps might introduce some unnecessary adjustments in certain cases. However, the significant drop in performance (-14.79%) when using strict span matching criteria highlights the challenge of exact boundary detection. This indicates that while our model is highly effective at identifying the general location of propaganda spans, pinpointing the exact character boundaries remains challenging.This in turn reflects how the model is unable to make the right predictions for the spans of those snippets marked not_propaganda.

Since our results show signs of possible overfitting when comparing to results in (Martino et al. 2020), we look in to the dataset which is of the shape (2414, 2) used for training and evaluating the model. We observe that 12 exact duplicate text samples were identified in the dataset which were labeled with either the same or different propaganda technique and over 305 pairs of near-duplicate texts with similarity greater then 0.9 were found. The presence of these duplicates across training and testing splits could artificially inflate performance metrics by allowing the model to memorize specific patterns rather than learn generalizable features. This raises important questions about the reported performance and whether it might reflect overfitting to duplicated examples rather than true generalization capability. Despite these concerns, the consistently high performance across different configurations and evaluation metrics suggests that our approach is effective for the propaganda span detection task.

With this approach our model was able to predict the spans of almost all propaganda technique types but failed with those sentences that were labeled not_propaganda. Comparing our results to the approach taken in (Chernyavskiy et al. 2020), we observe that predicting one span per sentence makes the problem less challenging when compared to predicting multiple spans in an article.

	Precision	Recall	F1-score	Support
flag_waving	0.6875	0.8462	0.7586	39
appeal_to_fear_prejudice	0.6047	0.6047	0.6047	43
causal_oversimplification	0.5385	0.6774	0.6000	31
doubt	0.4706	0.6316	0.5393	38
exaggeration,minimisation	0.5128	0.7143	0.5970	28
loaded_language	0.5455	0.4865	0.5143	37
name_calling,labeling	0.6923	0.5806	0.6316	31
repetition	0.3000	0.3750	0.3333	32
not_propaganda	0.9387	0.8140	0.8719	301
accuracy			0.7190	580
macro avg	0.5878	0.6367	0.6056	580
weighted avg	0.7509	0.7190	0.7295	580

Table 14: Classification Report - Using Spans generated with RoBERTa with BIO tagging

3.2.2 RoBERTa with Negative Sampling

The model's performance on the span identification task is summarized in Table 14, showing precision, recall, and F1-score metrics

Class	Precision	Recall	F1-score	Support
no_match	0.00	0.00	0.00	0
match	1.00	0.21	0.34	580
accuracy			0.21	580
macro avg	0.50	0.10	0.17	580
weighted avg	1.00	0.21	0.34	580

Table 15: Span Identification Performance Metrics

The results show perfect precision (1.00) for matched spans, indicating that when the model predicts a span, it is always correct. However, the recall is notably low (0.21), suggesting that the model fails to identify many propaganda spans in the test set. The resulting F1-score of 0.34 reflects this trade-off between high precision and low recall. The overall accuracy of 0.21 indicates that the model correctly identifies the exact propaganda span in approximately one-fifth of the test instances. The performance metrics reveal several important characteristics of the model:

1. The high precision indicates that the model is conservative in its predictions, only selecting spans when it has high confidence.

2. The low recall suggests that many propaganda spans are missed entirely, possibly due to:

   (a) Difficulty in identifying subtle propaganda techniques

   (b) Challenge in determining exact span boundaries

   (c) Complexity in capturing long-range contextual dependencies

The error analysis of our propaganda span detection model reveals several significant challenges. While achieving perfect precision (1.00), the model struggles with remarkably low recall (0.21), resulting in a modest F1-score of 0.34. This precision-recall imbalance indicates the model makes few false positive errors but frequently fails to identify actual propaganda spans. The primary issue involves boundary detection, where the model often identifies propaganda presence but fails to precisely capture the exact span boundaries. This manifests as partial span predictions, inclusion of extraneous words, or missing critical contextual elements—contributing significantly to the low recall since evaluation requires exact matches. Our 10-word limit on candidate spans, while computationally necessary, potentially handicaps detection of complex multi-phrase propaganda techniques, revealing a trade-off between computational efficiency and detection capability. Class imbalance emerges as another significant factor despite our negative sampling approach. Less frequent propaganda techniques receive inadequate representation in training data, hampering the model's ability to recognize these rarer categories. Our candidate-span classification strategy demonstrates significant promise through its perfect precision but requires substantial improvements to address the recall deficit.

	Precision	**Recall**	**F1-score**	**Support**
flag_waving	0.4839	0.7692	0.5941	39
appeal_to_fear_prejudice	0.4242	0.3256	0.3684	43
causal_oversimplification	0.5263	0.3226	0.4000	31
doubt	0.3415	0.3684	0.3544	38
exaggeration,minimisation	0.3043	0.5000	0.3784	28
loaded_language	0.2581	0.4324	0.3232	37
name_calling,labeling	0.3256	0.4516	0.3784	31
repetition	0.1220	0.3125	0.1754	32
not_propaganda	0.7760	0.4950	0.6045	301
accuracy			0.4672	580
macro avg	0.3958	0.4419	0.3974	580
weighted avg	0.5725	0.4672	0.4943	580

Table 16: Classification Report - Using Spans generated with RoBERTa with Negative Sampling

4 Discussion

Our exploration of multiple approaches for propaganda detection reveals significant performance variations across both tasks. For propaganda technique classification, RoBERTa-large achieved superior performance (macro F1=0.6694, micro F1=0.7812) compared to the SVM model (macro F1=0.2966) and the catastrophically underperforming T5 model (macro F1=0.0151). RoBERTa excelled at identifying techniques with clearer lexical patterns like ”flag waving” (F1=0.7955) and ”causal oversimplification” (F1=0.7059), leveraging its transformer architecture's ability to capture contextual relationships. The SVM's reliance on TF-IDF features limited its effectiveness with semantically complex categories, completely failing on ”loaded language” (F1=0.0000) while showing moderate success with lexically consistent techniques. The T5 model's reformulation of classification as text generation proved fundamentally unsuitable, degenerating into effectively a single-class predictor despite reasonable hyperparameter settings.

For span identification, the approaches demonstrated dramatically different behaviors. The BIO tagging method achieved remarkably high performance (F1=0.9914), though potential dataset artifacts warrant caution in interpreting these results. This approach directly models token-level classification, enabling precise boundary detection. The ablation study revealed that the model's raw predictions were optimal, with post-processing steps paradoxically reducing performance. Conversely, the negative sampling approach showed an extreme precision-recall imbalance (precision=1.00, recall=0.21, F1=0.34), demonstrating high confidence but failing to comprehensively identify propaganda spans. This approach's conceptualization of span identification as classification over candidate spans proved computationally intensive during inference and struggled with capturing the diversity of propaganda manifestations.

These performance disparities highlight important considerations for propaganda detection system design. Transformer-based approaches consistently outperformed traditional machine learning techniques, but architectural and methodological choices significantly impact effectiveness. The stark contrast between BIO tagging and negative sampling approaches for span identification demonstrates how task formulation fundamentally influences model behavior and practical utility. Furthermore, performance varied substantially across propaganda techniques, with models generally struggling on categories requiring broader contextual understanding or subtle semantic interpretation. These findings underscore the complexity of propaganda detection and suggest that different approaches may be optimal depending on specific application requirements and the nature of propaganda techniques being targeted.

5 Conclusion

This study presented multiple approaches for propaganda detection, addressing both technique classification and span identification tasks. Our findings demonstrate that transformer-based models, particularly RoBERTa, significantly outperform traditional machine learning approaches for propaganda detection, with the fine-tuned RoBERTa-large model achieving the highest performance (macro F1=0.6694) among our approaches for technique classification. The implementation of sequence labeling with BIO tagging proved exceptionally effective for span identification, achieving near-perfect performance (F1=0.9914), although we note that potential dataset artifacts warrant caution in interpreting these results. Our comparative analysis revealed that different propaganda techniques pose varying levels of challenge for automated detection. Techniques with clear lexical markers (like ”flag waving”) were more readily identified than those requiring broader contextual understanding (like ”repetition” and ”loaded language”). The stark performance gap between different approaches highlights the importance of model selection and training methodology, with the catastrophic failure of the T5-based approach highlighting that architectural sophistication does not guarantee effectiveness. The span identification task further demonstrated the inherent difficulty in precisely locating propaganda within text, with the negative sampling approach achieving perfect precision but struggling with recall. While significant progress has been made in propaganda detection, substantial challenges remain in developing models that can reliably identify and classify the full spectrum of propaganda techniques across diverse contexts, particularly when considering the real-world applications of such systems in combating misinformation.

6 Future Work

Future work could explore implementing Hierarchical Graph Interaction Networks (H-GIN) for propaganda detection, which have achieved state-of-the-art results with an F1 score of 0.80 (Ahmad et al. 2025). H-GIN's ability to model complex relationships between textual elements at different levels could be particularly effective for capturing the contextual nuances of propaganda techniques. We could also investigate more powerful language models like Flan-T5 as alternatives to our T5 implementation, building on (Sprenkamp et al. 2023) promising results with large language models for propaganda detection. For Task 2, while RoBERTa with BIO tagging showed excellent performance for single-span detection, we could explore hyperparameter optimization for the negative sampling approach to improve its recall while maintaining its perfect precision. Additionally, addressing class imbalance through advanced sampling techniques, exploring multimodal approaches that incorporate visual propaganda elements, and developing more robust evaluation methodologies could further advance propaganda detection capabilities.

References

Abdullah, Malak et al. (2022). "Detecting propaganda techniques in english news articles using pre-trained transformers". In: *2022 13th International Conference on Information and Communication Systems (ICICS)*. IEEE, pp. 301–308.

Ahmad, Pir Noman et al. (2025). "Hierarchical graph-based integration network for propaganda detection in textual news articles on social media". In: *Scientific Reports* 15.1, p. 1827.

Chernyavskiy, Anton et al. (2020). "Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning". In: *arXiv preprint arXiv:2008.02837*.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20, pp. 273–297.

Devlin, Jacob et al. (2019). "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: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Gururangan, Suchin et al. (2018). "Annotation artifacts in natural language inference data". In: *arXiv preprint arXiv:1803.02324*.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Martino, G et al. (2020). "SemEval-2020 task 11: Detection of propaganda techniques in news articles". In: *arXiv preprint arXiv:2009.02696*.

Raffel, Colin et al. (2020). “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *Journal of machine learning research* 21.140, pp. 1–67.

Sprenkamp, Kilian et al. (2023). “Large language models for propaganda detection”. In: *arXiv preprint arXiv:2310.06422*.