cocoa at SemEval-2025 Task 10: Prompting vs. Fine-Tuning: A Multilevel Approach to Propaganda Classification

Vineet Saravanan
Cranbrook Schools
Bloomfield Hills, MI
vineetsaravanan@gmail.com

Steven Wilson
University of Michigan-Flint
Flint, MI
steverw@umich.edu

Abstract

The increasing sophistication of natural language processing models has facilitated advancements in hierarchical text classification, particularly in the domain of propaganda detection. This paper presents our submission to SemEval 2025 Task 10, Subtask 1, which focuses on multilevel text classification for identifying and categorizing propaganda narratives in online news (Piskorski et al., 2025). We investigate two primary approaches: (1) prompt-based classification using large language models (LLMs) like GPT, which offers flexibility but struggles with hierarchical categorization, and (2) fine-tuning transformer-based models, where we employ a hierarchical structure—one model classifies the main propaganda category, followed by three separate models specializing in subcategory classification. Our results indicate that while LLMs demonstrate some generalization ability, fine-tuned models significantly outperform them in accuracy and reliability, reinforcing the importance of task-specific supervised learning for propaganda detection. Additionally, we discuss challenges related to data sparsity in subclassification and explore potential enhancements such as multi-task learning and hierarchical loss functions. Our findings contribute to the broader field of automated propaganda detection and emphasize the value of structured classification models in understanding patterns of online communication. All code and data used in our experiments will be made publicly available on our GitHub [1].

1 Introduction

The spread of propaganda and strategically crafted information in online media presents a growing challenge to public discourse and democratic processes. As propaganda techniques evolve, so too must the systems built to detect them. Traditional approaches to propaganda detection have largely focused on identifying specific rhetorical strategies—such as appeals to fear, doubt, or name-calling—at the sentence or span level (Da San Martino et al., 2019). Others have focused on classifying entire articles or highlighting binary instances of propagandistic content. While effective for technique recognition, these approaches fall short when propaganda is embedded through more subtle means, such as the strategic portrayal of specific named entities across a narrative.

Recent work has begun to shift toward understanding propaganda at the entity level, recognizing that how named entities are framed across a narrative can shape readers' perceptions. While earlier shared tasks and studies emphasized detecting rhetorical techniques at the sentence or span level, they did not require models to assess the narrative role of specific entities. The SemEval 2025 Task 10 builds on these foundations by introducing a more fine-grained challenge: Subtask 1 (Stefanovitch et al., 2025). In this task, systems are provided with a news article and a list of named entity (NE) mentions, and must assign one or more roles to each mention using a predefined taxonomy. These roles fall into three overarching categories—Protagonist, Antagonist, and Innocent—each with further fine-grained subtypes such as Saboteur or Conspirator, making this a multi-label, multi-class, span-level classification problem.

In this work, we explore two primary approaches to tackling this classification task: (1) prompting large language models (LLMs) such as GPT, leveraging their zero-shot capabilities for classification; and (2) fine-tuning transformer-based models, where we train one model for main category classification and three specialized models for subclassification within each category. While LLMs provide adaptability and require minimal task-specific training, our experiments indicate that fine-tuned models significantly outperform them in accuracy and reliability. This underscores the limitations of

[1] https://github.com/VSPuzzler/cocoa-at-SemEval-2025-Task-10

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 590–594
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

generic prompting in highly structured classification tasks and highlights the continued relevance of task-specific supervised learning.

By benchmarking different approaches to hierarchical propaganda detection, we aim to contribute insights that will inform future developments in automated media narrative analysis.

2 Related Work

Early efforts in propaganda detection focused primarily on identifying rhetorical techniques within news articles. For example, SemEval-2020 Task 11 introduced two subtasks: Span Identification and Technique Classification. These tasks aimed to detect specific propaganda techniques, such as "loaded language" or "appeal to fear", within textual spans, using models such as BERT for improved accuracy (Martino et al., 2020).

Building upon this foundation, recent studies have explored the capabilities of large language models (LLMs) in propaganda detection. An investigation was carried out on the performance of GPT-4 in identifying propagandist content. The findings indicated that, while LLMs show promise, they often perform underperformance compared to fine-tuned models, especially in tasks that require a nuanced understanding of context and subtle linguistic cues (Szwoch et al., 2024).

Advancing the field further, there was an introduction of a multilingual hierarchical corpus specifically designed for entity framing and role portrayal in news articles. The dataset categorizes entities into fine-grained roles nested within three main categories: protagonist, antagonist, and innocent. This taxonomy facilitates a more detailed analysis of how entities are portrayed in different narratives and languages (Mahmoud et al., 2025).

3 Methodology

Our approach involved a structured pipeline comprising three main stages: data preprocessing, LLM-based classification using GPT, and fine-tuning a transformer-based model for hierarchical classification. Each stage was designed to efficiently extract entity roles from news articles and classify them into Protagonist, Antagonist, or Innocent, along with their respective subcategories.

3.1 Data Processing

To prepare the dataset for classification, we first extracted entity role annotations from the provided subtask-1-annotations.txt file combining batches 1-3. We only looked at the English documents for this study. Each annotation included a document filename, entity role, and character position within the text. A preprocessing script parsed this information, identifying the main category (Protagonist, Antagonist, Innocent) and grouping any corresponding subcategories. To match the annotations with the corresponding news articles, we loaded and indexed all raw documents from the dataset directory. This ensured each document was correctly paired with its respective annotations.

3.2 Zero-Shot GPT-Based Prompting

We first experimented with LLM-based prompting using different GPT models. This approach required formulating structured prompts to guide the model through a two-tier classification process. The first step involved main category classification, where the entire article was provided to GPT, followed by the question: "Is this article about a Protagonist, Antagonist, or Innocent?" To determine the predicted category, we compared the model's response with each of the three possible labels using a similarity function based on the SequenceMatcher algorithm. The category with the highest similarity score was selected as the predicted main category. Tests were done with and without the target word in the prompt.

Once a main category was assigned, a follow-up subclassification prompt was generated. Each prompt listed the potential subcategories relevant to the assigned category and instructed GPT to choose one or more applicable labels. For instance, if the main category was classified as Protagonist, GPT was asked to select from Guardian, Martyr, Peacemaker, Rebel, Underdog, and Virtuous. Similarly, tailored prompts were used for Antagonist (e.g., Instigator, Conspirator, Tyrant, Saboteur, etc.) and Innocent (e.g., Forgotten, Exploited, Victim, Scapegoat). The model's responses were parsed, and subcategories were assigned based on keyword matching.

Despite the flexibility and zero-shot capabilities of GPT-based classification, this method exhibited lower accuracy due to inconsistencies in response formatting and difficulties in handling hierarchical label dependencies. This motivated our shift towards fine-tuning a transformer model for more precise classification.

3.3 Fine-Tuning Transformer-Based Models

To improve classification accuracy, we fine-tuned a BERT-based transformer model for hierarchical classification. We selected BERT-base-uncased as the base model due to its strong performance in text classification tasks. To enhance its ability to detect entity roles, we modified the tokenizer by introducing two special tokens—`[TARGET]` and `[/TARGET]`—which explicitly marked the span of interest within the text. This ensured that the model focused on the relevant entity while still considering the surrounding context. To preprocess the text, we inserted these special tokens at the entity's start and end positions based on the provided annotations. Since the model has a 512-token limit, we applied a context-aware truncation strategy, prioritizing the region around the marked entity to retain as much relevant information as possible.

The dataset was tokenized using Hugging Face's AutoTokenizer and split into training and test sets (80/20 split randomly), ensuring a balanced distribution of the three main categories and their subcategories. Each data instance was encoded to include input IDs, attention masks, and corresponding labels for both the main category and subcategories. A custom PyTorch dataset class was implemented to facilitate structured data loading for training. We leveraged the Hugging Face Trainer API, setting hyperparameters including a learning rate of 2e-5, batch size of 8, weight decay of 0.01, and a total of 3 training epochs.

To evaluate the model, we computed classification accuracy for both the main category and subclassification tasks. The model's final weights and tokenizer were saved and uploaded to the Hugging Face Model Hub for reproducibility and potential future improvements. Compared to the GPT-based prompting approach, this fine-tuned model demonstrated superior accuracy and consistency, reinforcing the benefits of task-specific supervised learning for structured propaganda classification.

4 Results

To compare the performance of GPT-4o, GPT-3.5-turbo, and GPT-3.5-turbo-1106, we evaluated each model's ability to classify news articles into the main categories (Protagonist, Antagonist, Innocent) and their corresponding subcategories. The accuracy of each model was calculated based on its ability to correctly assign labels to a test set using prompt-based classification. Initially, prompts that included a highlighted target word resulted in 0% accuracy across all models. Subsequent tests removed the explicit target word, which led to modest improvements in performance.

Model	Main Category Accuracy	Subcategory Accuracy
GPT-4o	19.10%	0.00%
GPT-3.5-turbo	23.47%	0.00%
GPT-3.5-turbo-1106	22.16%	0.00%

Table 1: Comparison of GPT Models for Main and Subcategory Classification Without Target Word

From the results, GPT-3.5-turbo achieved the highest main category accuracy (23.47%), outperforming GPT-4o (19.10%) and GPT-3.5-turbo-1106 (22.16%). However, all GPT models failed to classify subcategories correctly, with an accuracy of 0% across all models. This indicates that while GPT models can somewhat differentiate between Protagonist, Antagonist, and Innocent roles, they struggle with fine-grained subclassification, likely due to the lack of explicit hierarchical dependencies in their zero-shot and few-shot prompting approach. Additionally, prompt structure played a critical role. Slight phrasing changes led to significant shifts in predictions, suggesting that model performance is highly sensitive to instruction design. For instance, placing the named entity mention at different positions in the prompt sometimes caused role confusion. These results highlight the limitations of LLM-based classification for structured multi-level tasks, where fine-tuned models may be necessary to achieve reliable subclassification performance.

To address the limitations of LLM prompting, we fine-tuned multiple transformer-based models, including BERT-base (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and ELECTRA (Clark et al., 2020). These models were trained using the tokenized dataset with entity span markers, and their performance was evaluated based on accuracy for both main category classification and subclassification. The results are summarized in Tables 2 and 3 using equations 1-3.

Model	Main Category Accuracy	Protagonist Accuracy, F1
bert-large-uncased	68.11%	91.03%, 0.00
roberta-base	68.11%	91.00%, 0.00
distilbert-base-uncased	68.11%	91.03%, 0.00
google/electra-base-discriminator	68.11%	91.03%, 0.00

Table 2: Main Category Accuracy and Protagonist Performance

Model	Antagonist Accuracy, F1	Innocent Accuracy, F1
bert-large-uncased	90.89%, 0.00	78.10%, 0.56
roberta-base	90.90%, 0.00	78.10%, 0.56
distilbert-base-uncased	90.89%, 0.00	78.10%, 0.56
google/electra-base-discriminator	90.89%, 0.00	78.10%, 0.56

Table 3: Subcategory Performance: Antagonist and Innocent Roles

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

The results indicate that all fine-tuned transformer models—BERT-Large, RoBERTa, DistilBERT, and ELECTRA—achieved identical main category accuracy (68.11%), suggesting that model architecture had little impact on distinguishing between the categories. Additionally, subcategory classification results remain nearly unchanged across models, with Protagonist accuracy around 91.03%, Antagonist at 90.89%, and Innocent at 78.10%, while F1 scores for Protagonist and Antagonist remain at 0.00.

This uniformity in results suggests potential limitations in the dataset and label distribution, as fine-tuned models typically exhibit more variation in classification tasks. The lack of improvement in F1 scores, particularly for the Protagonist and Antagonist categories, indicates that while the models may identify relatively correct probabilities for each subcategory, they are not able to predict each one exactly. This may be because in many cases, there can be multiple subclassifications. Furthermore, the consistent 78.10% accuracy and 0.56 F1 score for Innocent classification suggest that this category may have a more balanced/better-defined representation in the dataset than the others.

Overall, fine-tuned transformers outperformed GPT models in both main category and subcategory classification, demonstrating the importance of task-specific supervised learning. While GPT models provided rapid inference without additional training, they exhibited inconsistencies in subcategory assignments and required manual prompt engineering to improve reliability. In contrast, fine-tuned models provided more stable predictions, particularly RoBERTa and BERT-base, which effectively leveraged hierarchical label structures to improve classification accuracy.

These findings suggest that while LLMs can serve as an initial baseline, fine-tuned transformer models remain the preferred approach for hierarchical classification tasks, particularly when detailed label hierarchies are involved. Future improvements could involve hybrid approaches, integrating the generalization capabilities of LLMs with the structured learning of fine-tuned models to further enhance classification performance.

The results of the final submission is in Table 4.

Rank	Exact Match Ratio	micro P	micro R	micro F1	Accuracy for main role
30	0.01700	0.08750	0.12450	0.10280	0.82550

Table 4: Performance metrics for team "cocoa" on SemEval Task 10 Subtask 1

While our model demonstrated strong performance in predicting the main role, achieving high accuracy, the lower exact match ratio and F1 score indicate challenges in correctly predicting both the main category and subcategories simultaneously. These results suggest that while our fine-tuned models effectively captured broad entity roles, subclassification remains a difficult task, likely due to data sparsity and overlap between subcategories. Future improvements could focus on better handling hierarchical dependencies, leveraging external knowledge sources, or adopting contrastive learning techniques to refine subcategory classification.

5 Conclusion

In this work, we explored two approaches for hierarchical propaganda classification in SemEval 2025 Task 10, Subtask 1: LLM-based prompting and fine-tuned transformer models. Our results demonstrated that while GPT models offered a flexible, zero-shot solution, they struggled with hierarchical dependencies and inconsistent subcategory classification. In contrast, fine-tuned transformer models, particularly RoBERTa and BERT-base, significantly outperformed LLMs, achieving higher accuracy for both main category and subcategory classification.

These findings highlight the importance of task-specific supervised learning in structured classification tasks. Future work could explore hybrid approaches that combine LLMs for generalization with fine-tuned models for precision. Additionally, integrating external knowledge sources or multi-task learning frameworks could further improve classification accuracy. Our study contributes to the development of automated propaganda analysis and provides a foundation for more advanced methods in entity framing detection within news media.

References

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. 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)*, pages 5636–5646. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, et al. 2025. Entity framing and role portrayal in the news. *arXiv preprint arXiv:2502.14718*.

G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, PurificaÃ§Ã£o Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo GuimarÃ£es, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. 2024. Limitations of large language models in propaganda detection task. *Applied Sciences*, 14(10):4330.