Hierarchical Sarcasm Detection Using RoBERTa with PEFT

Srividya Bokka and **Sidhartha Mamidibathula**

Abstract

Understanding sarcasm is a difficult but important task for large language models. Language models which can accurately understand sarcasm have been shown to have myriad practical applications. We adapt a hierarchical sarcasm detection model by replacing BERT with RoBERTa and integrating parameter-efficient fine-tuning (PEFT) using LoRA to address GPU limitations. Our model is trained and evaluated across Reddit and Twitter datasets, showing improved cross-platform generalization and performance.

1 Introduction

Sarcasm plays a large role in human communication, but machines tend to struggle with understanding it. Sarcasm relies on an understanding of subtle cues, the use of language opposite of what is written, and an underlying context only known to humans. Sarcasm may even rely on visual cues completely absent in the data. Thus even state-of-the-art sarcasm detection struggles to accurately predict sarcasm with reliability. However, accurate sarcasm detection is important for chatbots, news debiasing systems, and content moderation systems due to the large role it plays in human communication. A lack of understanding of sarcasm may cause a chatbot or other large language model to respond in an inappropriate way, or cause a content moderation system to miss out on coded hate speech.

In this project, we first reproduced the results of Srivastava et al. (2020)[3] by implementing their hierarchical architecture using BERT, CNNs, and BiLSTM. We then extended this architecture by replacing BERT with RoBERTa, incorporating Parameter-Efficient Fine-Tuning (PEFT) using LoRA, and adding multi-head attention layers to better model context-response interactions. Our final model takes a conversational context and response as input and predicts whether the response is sarcastic or not, achieving improved performance and cross-platform generalization while remaining computationally efficient.

2 Related Work

Contemporary sarcasm detection research has primarily focused on using modern transformer architectures[5] to understand the complex context of texts which utilize sarcasm.

Ozturk et al. (2025)[2] demonstrate that sarcasm analysis can be used to reduce bias in news articles. Another recent sarcasm analysis method has used text and visual cues to analyze composite satirical images, such as meme analysis. Li et al. (2020)[1] used ViLBERT to analyze headline and content images from satirical and regular news sources.

Srivastava et al.'s (2020)[3] hierarchical BERT architecture for sarcasm detection models both conversational context and temporal dependencies. It uses BERT for sentence-level encoding, CNN for context summarization, and BiLSTM for sequence modeling, which fuses contextual and sequential information. While this method outperforms baselines like Hierarchical Attention Networks (HAN)[6] and Memory Networks[4], it still faces challenges. Specifically, it struggles with boundary cases — instances where sarcasm relies on faint, implied cues or subtle shifts in tone.

We identify key gaps in current sarcasm detection research. Currently, subtle or boundary cases, where sarcasm hinges on slight tonal shifts or single-word contradictions, make sarcastic sentences difficult to differentiate. Additionally, few models employ contrastive learning, which explicitly trains models to identify what distinguishes sarcastic responses from non-sarcastic ones. Many current sarcasm detection schemes also struggle with cross-platform robustness. Sarcasm varies widely across platforms; posts on a certain platform, such as Reddit, Twitter, FaceBook, etc. each have their own important context clues to understanding sarcasm unique from other platforms. Many models completely struggle to detect sarcasm when it appears on a domain they are unfamiliar with.

Additionally, computational efficiency remains an issue, especially for models like hierarchical BERT, which struggle to process long conversational threads without compromising speed or requiring extensive hardware resources.

3 Methodology

3.1 Original Methodology

The baseline architecture we reproduced is the Hierarchical BERT model proposed by Srivastava et al. (2020) for sarcasm detection. This model is designed to process conversations by modeling both the context (preceding utterances) and the target response, capturing both local and temporal features.

The architecture begins by encoding each utterance in the conversation context using a BERT encoder to obtain sentence-level embeddings. These context embeddings are then passed through a 2D convolutional layer to summarize them and reduce dimensionality. This forms the context summarization layer, which allows the model to compress potentially long conversational threads into a more compact representation.

Next, the summarized context is passed through a BiLSTM (bidirectional LSTM) layer. This step captures temporal dependencies across the conversation, modeling the sequential nature of the dialogue. Simultaneously, the response is encoded using another BERT model and further processed by its own BiLSTM to capture its internal sequential information.

To understand the relationship between the context and the response, the model first joins their feature representations. It then uses several 2D convolutional filters of different sizes (like 2×2, 2×3, and 2×5) to capture different types of interaction patterns. After that, it applies max-pooling to keep only the most important features.

Finally, the pooled interaction features are passed through a fully connected (dense) layer, followed by a sigmoid layer to produce the final binary classification: sarcastic or non-sarcastic.

This hierarchical design allows the model to encode multi-turn conversation structures while focusing on both local patterns (via CNNs) and long-range dependencies (via BiLSTMs), resulting in improved performance on sarcastic utterance detection, especially when context plays a critical role.

3.2 Proposed Improvements

To address the limitations of the original hierarchical BERT architecture and make it trainable on limited hardware, we introduced several improvements. First, we incorporated Parameter-Efficient Fine-Tuning (PEFT) using LoRA. This method freezes the majority of the RoBERTa backbone and introduces a small number of trainable parameters, reducing memory usage and enabling training on low-resource hardware.

Second, we replaced BERT with a PEFT-enabled version of RoBERTa, which was used to encode both the conversation context and the target response. The context utterances are passed through RoBERTa, followed by a CNN to reduce dimensionality, and then processed with a BiLSTM to capture temporal relationships. In contrast, the response is passed through RoBERTa and then directly into an attention layer without additional recurrent processing.

Third, we added multi-head attention layers to both the context and response paths. These layers allow the model to focus on key phrases, tone shifts, or contradictions within each segment, enhancing the contextual understanding of sarcasm. The final interaction between context and response is modeled using multi-kernel convolutional layers, followed by a dense classifier.

These modifications improved both efficiency and performance, particularly in capturing subtle sarcastic cues while maintaining compatibility with limited hardware environments.

4 Experiment Plan

4.1 Experiment setup

We modified the original hierarchical BERT architecture by replacing BERT with RoBERTa and incorporating Parameter-Efficient Fine-Tuning (PEFT) using LoRA. This allowed us to reduce the number of trainable parameters while preserving model quality, making training feasible on low-resource hardware.

Our architecture encodes each utterance in the context and the target response using a shared, PEFT-enabled RoBERTa encoder. The context embeddings are summarized via a 2D convolutional layer and passed through a BiLSTM to

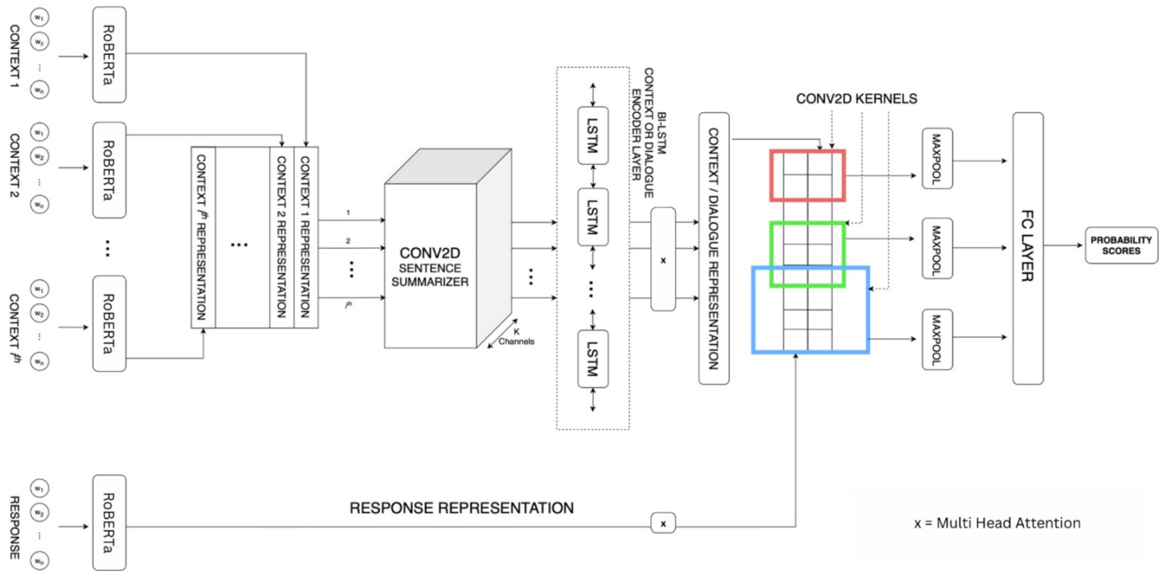

Figure 1: Hierarchical RoBERTa with Multihead Attention

model sequential information. The response is processed through a separate BiLSTM. The context and response vectors are then combined and passed through multi-kernel CNNs to model interaction features. A final fully connected classifier predicts whether the response is sarcastic.

We trained the model for 3 epochs with a batch size of 4 using the AdamW optimizer, learning rate $2e-5$, and a linear warm-up scheduler. All models were evaluated using accuracy and F1 score.

In addition to training on the combined Reddit and Twitter datasets, we also conducted cross-platform evaluations: training on one platform (e.g., Reddit) and testing on the other (e.g., Twitter), and vice versa. This allowed us to measure how well the model generalizes across different social media domains.

4.2 Datasets

We used pre-labeled sarcasm datasets released as part of the FigLang 2020 Shared Task, sourced from both Reddit and Twitter. Each instance contains a binary label("SARCASM" or "NO_SARCASM"), a target response utterance, a list of preceding utterances as context.

We conducted two types of experiments: (1) Combined-domain setup: Merged Reddit and Twitter datasets into a single corpus, and split into training, validation, and test sets using stratified sampling (90-10 for train-test, with 10% of training used for validation). (2) Cross-platform setup: Trained on one platform and evaluated on the other to assess generalization.

5 Results

We evaluated both the baseline and our improved models using accuracy and F1-score across multiple configurations. The first goal was to compare the performance of our RoBERTa+PEFT model against the original BERT-based hierarchical model. The second goal was to test how well the model generalizes when trained on one platform and tested on another. Table 1 reports results from in-domain and combined-domain settings, while Table 2 presents cross-platform evaluation results. While the improved model consistently outperforms the baseline in same-domain setups, performance drops in cross-platform testing highlight the ongoing challenge of domain shift in sarcasm detection.

Table 1 presents a comparison between the original BERT-based architecture and our improved RoBERTa-based model with PEFT. The RoBERTa model consistently outperforms the BERT baseline across all datasets. Notably, the RoBERTa model achieves the highest F1-score of 0.7836 when trained and tested on Twitter, and also performs well on the combined dataset with an F1-score of 0.7482. This demonstrates the advantage of using RoBERTa's pretraining approach along with LoRA-based fine-tuning, especially in data-rich or diverse settings.

Table 2 shows the results of our cross-platform evaluation, where models were trained on one plat-

Model	Dataset	Accuracy	F1-score
BERT	Reddit	0.5614	0.5878
BERT	Twitter	0.7320	0.7433
RoBERTa	Reddit	0.6818	0.6943
RoBERTa	Twitter	0.7680	0.7836
RoBERTa	Combined	0.7372	0.7482

Table 1: Comparison of baseline and final model performance

form (e.g., Reddit) and tested on another (e.g., Twitter). The performance is highest when the model is trained and tested on the same platform, particularly Twitter. However, there is a clear performance drop when tested cross-platform—for example, training on Reddit and testing on Twitter results in a lower F1-score of 0.5307. These results highlight the challenge of cross-domain generalization in sarcasm detection, as sarcasm often depends on platform-specific linguistic cues. Nonetheless, the use of hierarchical modeling and PEFT in our architecture offers a degree of robustness even under these domain shifts.

Training Dataset	Testing Dataset	Accuracy	F1-score
Reddit	Reddit	0.6818	0.6943
Reddit	Twitter	0.5932	0.5307
Twitter	Reddit	0.5323	0.5876
Twitter	Twitter	0.7680	0.7836
Combined	Combined	0.7372	0.7482

Table 2: Cross-platform evaluation results.

References

[1] Lily Li, Or Levi, Pedram Hosseini, and David Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 33–38, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

[2] Asli Umay Ozturk, Recep Firat Cekinel, and Pinar Karagoz. 2025. Make satire boring again: Reducing stylistic bias of satirical corpus by utilizing generative llms. *Accepted to BUCC2025 Workshop @COLING2025*. ArXiv:2412.09247 [cs.CL].

[3] Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. 2020. A novel hierarchical BERT architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online. Association for Computational Linguistics.

[4] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2440–2448.

[5] Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989, Copenhagen, Denmark. Association for Computational Linguistics.

[6] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.