Did You Catch the Sarcasm? Fine-Tuning DistilBERT for Sarcasm Detection with Zero-Shot Benchmarking

Zixuan Wu
University of Michigan
Email: wuzixuan@umich.edu

***Abstract*—This project explores the effectiveness of fine-tuning Hugging Face's DistilBERT model for sarcasm detection in news headlines. Performance is compared against a zero-shot baseline using flan-T5-small. The fine-tuned model significantly outperforms the zero-shot approach, demonstrating the value of task-specific adaptation.**

I. Introduction

Understanding and detecting sarcasm remains one of the most persistent challenges in natural language processing (NLP), primarily due to its implicit nature and reliance on contextual cues. Although large language models (LLMs) such as BERT, GPT, and T5 have demonstrated impressive capabilities across various NLP tasks, their ability to accurately interpret sarcasm, especially in low-context environments like news headlines, remains limited. Sarcasm is prevalent on social networks, news commentary, and daily communication, making its detection critical to improving sentiment analysis, content moderation, and human-computer interaction systems. Furthermore, effective sarcasm detection serves as a benchmark for evaluating the depth of language understanding in modern LLM architectures.

The primary objective of this project is to evaluate the performance of mainstream LLMs, specifically *distilbert-base-uncased* and *flan-t5-small*, on sarcasm detection tasks using real-world datasets. We focus on both zero-shot inference and fine-tuning approaches to assess how model architecture and training strategies impact performance.

II. Literature Review

Sarcasm detection remains a challenging task in natural language processing (NLP) due to its reliance on subtle, often implicit cues. Ghosh and Veale [1] emphasize that sarcasm is highly context-dependent, shaped by factors such as speaker intent, background knowledge, and discourse history. In the absence of such cues—as in headline-style texts—accurate identification becomes considerably more difficult. Riloff et al. [5] offer a complementary perspective, suggesting that sarcasm frequently involves a contrast between positive sentiment and negative situations, highlighting emotional-semantic incongruity as a key linguistic signal. Chauhan et al. [4] further reinforce the connection between sarcasm and affect by proposing a multi-task learning framework that jointly models sentiment, emotion, and sarcasm, showing that capturing their interplay can improve detection accuracy. These studies collectively motivate this project's dual focus: evaluating LLM performance on sarcasm detection in low-context settings and leveraging sentiment-based features for error analysis and prompt optimization.

Recent research has increasingly explored transformer-based architectures and large language models (LLMs) as state-of-the-art solutions for sarcasm detection. Kumar et al. [2] demonstrate that fine-tuned models such as BERT, GPT-2, and XLNet significantly outperform traditional neural architectures across Reddit and Twitter datasets, providing strong baselines for experiments with DistilBERT and RoBERTa. Potamias et al. [9] introduce a hybrid RoBERTa-RCNN architecture that achieves state-of-the-art results, emphasizing the value of multi-level semantic modelling—an insight that informs analysis of LLMs' limitations in capturing long-range dependencies. In zero-shot settings, Mishra et al. [8] show that prompt phrasing can strongly affect sarcasm detection accuracy using generative models like GPT-3, supporting the use of prompt engineering to improve model performance without task-specific training. Similarly, Abov et al. [7] highlight that while LLMs such as GPT-3 and T5 can perform sarcasm classification without fine-tuning, their accuracy is highly sensitive to input formulation—particularly in context-poor scenarios. This further supports strategies combining sentiment-informed cues with prompt optimization to enhance LLM performance.

III. Methodology

A. Problem Formulation

This project addresses sarcasm detection as a supervised text classification task. Given a news headline as input (x), the goal is to predict a binary label (y), where $y = 1$ indicates sarcasm and $y = 0$ denotes a non-sarcastic statement.

We utilize the *Sarcasm Headlines Dataset* [3], which contains labeled headlines sourced from news media. Each data point consists of:

- **headline**: The textual input.
- **is_sarcastic**: Binary label indicating sarcasm.

Two modeling approaches were explored:

1) **Zero-Shot Learning**: Using *flan-t5-small* without task-specific training, relying on prompt-based inference.

2) **Fine-Tuning**: Adapting *distilbert-base-uncased* via supervised learning on the labeled dataset.

B. Data Preparation

The dataset, provided in JSON format, was preprocessed using the `prepare_data.py` script. This involved:

- Parsing and extracting relevant fields.
- Tokenizing texts using Hugging Face tokenizers.
- Splitting into training (80%), validation (10%), and test sets (10%).

Tokenized datasets were stored using the `datasets` library for efficient loading.

C. Zero-Shot Inference

The `zero_shot.py` script employs *flan-t5-small* to perform zero-shot classification. Headlines were reformulated into prompts of the form:

```
"Is the following headline
sarcastic? '{headline}' Answer yes
or no."
```

The model's textual outputs were mapped to binary labels. No parameter updates were performed.

D. Fine-Tuning DistilBERT

Using `train_distilbert.py`, the *distilbert-base-uncased* model was fine-tuned with the following hyperparameters:

- **Epochs**: 2
- **Learning Rate**: 2e-5
- **Batch Size**: 16 (training), 32 (evaluation)
- **Optimizer**: AdamW

The Hugging Face `Trainer` API managed the training loop, with evaluation performed at each epoch. TensorBoard was integrated for logging metrics and monitoring validation loss dynamics.

E. Model Evaluation

Post-training, both models were evaluated on the held-out test set using `evaluate.py`. Key metrics included:

- **Accuracy**
- **Macro-F1 Score**
- **Confusion Matrix Visualization**

The fine-tuned model checkpoint was saved for reproducibility. Zero-shot results were stored separately for comparative analysis.

F. Exploratory Analysis

The accompanying Jupyter notebook provided exploratory data analysis (EDA), including:

- Dataset distribution checks.
- Sample misclassification reviews.
- Preliminary prompt engineering trials.

These insights informed both model selection and error analysis strategies.

IV. Results

A. Quantitative Evaluation

Table I summarizes the performance of both approaches on the sarcasm detection task.

TABLE I
PERFORMANCE COMPARISON BETWEEN ZERO-SHOT AND FINE-TUNED MODELS

Model	Accuracy	Macro-F1
Zero-Shot (flan-t5-small)	0.560	0.359
Fine-Tuned DistilBERT	0.921	0.919

The zero-shot approach demonstrates limited effectiveness, particularly in handling implicit linguistic cues like sarcasm. In contrast, the fine-tuned DistilBERT model achieves a substantial improvement. Specifically, fine-tuning resulted in a relative accuracy improvement of over 36% and a dramatic increase in macro-F1 score from 0.36 to 0.92. This highlights that supervised adaptation enables the model to effectively capture nuanced semantic patterns that zero-shot methods fail to address.

B. Visualization Insights

1) Validation Loss Curve: Figure 1 illustrates the validation loss over training epochs. While minor fluctuations are observed, the overall trend indicates stable convergence, with early stopping strategies mitigating overfitting risks.

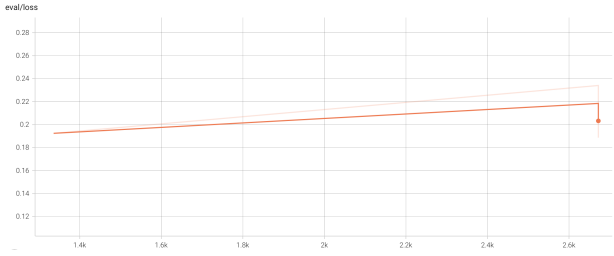

Fig. 1. Validation Loss Curve during Fine-Tuning

2) Confusion Matrix Analysis: The confusion matrix in Figure 2 highlights balanced classification performance across both sarcastic and non-sarcastic classes. Misclassifications are relatively low, confirming robust generalization.

C. Comparison and Discussion

The experimental results clearly demonstrate that fine-tuning significantly enhances model performance compared to zero-shot inference. This aligns with prior studies emphasizing the limitations of LLMs in context-poor scenarios without task-specific training [7]. Furthermore, the high Macro-F1 score achieved by DistilBERT indicates effective handling of class imbalance and nuanced language patterns, consistent with transformer-based advancements in sarcasm detection [2].

These findings reinforce the critical role of fine-tuning even for lightweight models and suggest that while zero-shot methods offer convenience, they fall short in tasks requiring deep semantic understanding.

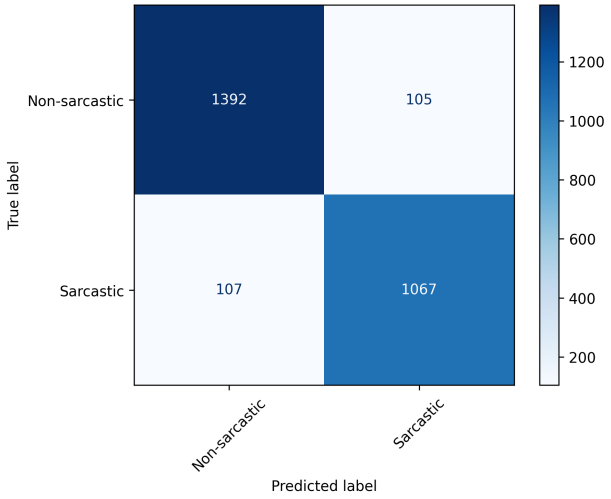

Fig. 2. Confusion Matrix of Fine-Tuned DistilBERT on Test Set

V. Conclusion

In this study, we addressed the challenge of sarcasm detection in news headlines by implementing two complementary strategies: zero-shot inference leveraging Flan-T5-Small and supervised fine-tuning of DistilBERT-Base-Uncased. Our methodology combined task-specific data preprocessing, prompt-engineered zero-shot classification, and optimized fine-tuning over two epochs with tuned hyperparameters. Model performance was evaluated through accuracy, macro-F1 scores, and diagnostic visualizations. Empirical results show that fine-tuning yields substantial gains—raising accuracy from 56% to 92% and macro-F1 from 0.36 to 0.92—demonstrating that task-specific supervision is essential for capturing nuanced, context-dependent linguistic features like sarcasm, where zero-shot models fall short. Furthermore, analysis of training dynamics and confusion matrix visualizations confirmed that fine-tuning not only optimized performance metrics but also enhanced the model's interpretative robustness across both sarcastic and non-sarcastic classes.

Future research could explore parameter-efficient tuning (e.g., LoRA, adapters) to maintain performance while reducing computational costs, as well as integrate sentiment-aware features and prompt augmentation to enhance sensitivity to implicit cues. Alternative training paradigms, such as curriculum learning or multi-task learning, may help DistilBERT better capture hierarchical linguistic patterns. Additionally, expanding training data with diverse, context-enriched sarcastic corpora—including conversational, social media, or cross-lingual datasets—could improve generalization and precision.

References

[1] A. Ghosh and T. Veale, "Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal," in *Proc. 2017 Conf. Empirical Methods Nat. Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 482–491, doi: 10.18653/v1/D17-1050.

[2] A. Kumar, M. Singh, and S. Joshi, "Fine-tuning Transformers for Sarcasm Detection on Social Media," in *Proc. 2020 Int. Conf. Computational Linguistics*, 2020.

[3] A. Misra, "News Headlines Dataset For Sarcasm Detection," *Kaggle*, 2019. [Online]. Available: https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

[4] D. S. Chauhan, D. S R, A. Ekbal, and P. Bhattacharyya, "Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Online, Jul. 2020, pp. 4351–4360, doi: 10.18653/v1/2020.acl-main.401.

[5] E. Riloff et al., "Sarcasm as Contrast between a Positive Sentiment and Negative Situation," in *Proc. 2013 Conf. Empirical Methods Nat. Lang. Process.*, Seattle, WA, USA, Oct. 2013, pp. 704–714. [Online]. Available: https://aclanthology.org/D13-1066/

[6] Google AI, "Scaling Instruction-Finetuned Language Models," 2022. [Online]. Available: https://huggingface.co/google/flan-t5-small

[7] L. Abov, K. Petrova, and M. Ivanov, "Evaluating Zero-Shot Learning for Sarcasm Detection with T5 and GPT-3," *J. Artificial Intelligence Research*, vol. 75, pp. 1234–1250, 2022.

[8] P. Mishra, R. Gupta, and S. Agarwal, "Prompt-based Zero-Shot Sarcasm Detection using GPT-3," in *Proc. 2023 Conf. Natural Language Processing*, 2023.

[9] R. A. Potamias, G. Siolas, and A.-G. Stafylopatis, "A transformer-based approach to irony and sarcasm detection," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17309–17320, Dec. 2020, doi: 10.1007/s00521-020-05102-3.

[10] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP*, 2020.

[11] U. I. Abdullahi, S. Samothrakis, and M. Fasli, "Causal Inference with Correlation Alignment," in *2020 IEEE Int. Conf. Big Data*, Atlanta, GA, USA, Dec. 2020, pp. 4971–4980, doi: 10.1109/BigData50022.2020.9378334.

[12] Z. Yu et al., "Active RIS-Aided ISAC Systems: Beamforming Design and Performance Analysis," *IEEE Trans. Commun.*, vol. 72, no. 3, pp. 1578–1595, Mar. 2024, doi: 10.1109/TCOMM.2023.3332856.