Read between the Headlines:
NYU MSCS Deep Learning – Final Project

Khwaab Thareja, Rachit Pathak, Xiaozhou Wen
Department of Computer Science, New York University
{kt3180, rmp10015, xw3795}@nyu.edu

1 Problem Statement

Sarcasm detection remains challenging due to implicit meaning, world knowledge, and pragmatic cues. Errors here affect downstream NLP systems such as summarization, sentiment analysis, and misinformation detection.
This project investigates whether a large instruction-tuned LLM (LLaMA 3) fine-tuned via parameter-efficient methods (LoRA) can outperform a traditional transformer encoder (BERT) on binary headline-level sarcasm detection.
We hypothesize that fine-tuning a Large Language Model (Llama 3-8B), which possesses extensive pre-trained world knowledge, will outperform a fine-tuned encoder-only model (BERT) in detecting sarcasm, particularly in headlines requiring external context.

Keywords: Deep Learning, NLP, Sarcasm Detection, BERT, Llama 3, Fine-tuning, LoRA.

2 Dataset Description

The dataset used is the News Headlines Dataset for Sarcasm Detection, sourced from Kaggle. It is collected from two news websites: TheOnion.com (sarcastic headlines) and HuffingtonPost.com (non-sarcastic headlines), and is licensed under standard Kaggle terms (likely CC BY-SA). The dataset contains approximately 28,000 samples in JSON format, with each sample consisting of a headline (text string input), an `is_sarcastic` label (binary: 1 for sarcastic, 0 for non-sarcastic), and an article link. Roughly 13,000 samples are sarcastic.
Preprocessing will include tokenization, lowercasing, removal of special characters, and data augmentation techniques like synonym replacement to handle class imbalance if needed. We will use an 80/10/10 train/validation/test split. Potential biases include source-specific stylistic differences (e.g., TheOnion's satirical tone vs. HuffPost's factual style), which may affect generalization to other domains, and possible cultural or temporal biases in the headlines collected.

3 Proposed Model and Technical Approach

We will compare two transformer architectures: BERT-base-uncased and LLaMA 3-8B. Both models will be fine-tuned rather than trained from scratch due to dataset size constraints and transfer learning benefits. BERT's bidirectional attention is well-suited for contextual understanding, while LLaMA 3's larger capacity may better capture subtle linguistic patterns in sarcasm.
For BERT, we'll add a classification head on the [CLS] token embedding. For LLaMA 3, we'll extract the last token embedding and add a linear classifier. Both models will use binary cross-entropy loss with logits. Optimization will employ AdamW with learning rates of 2e-5 (BERT) and 5e-6 (LLaMA 3), with linear warmup and cosine decay. Regularization includes dropout (0.1 for BERT, 0.2 for LLaMA 3), weight decay (0.01), and gradient clipping (max norm 1.0).

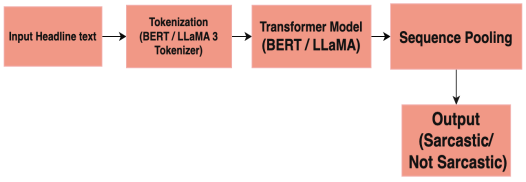

Figure 1: Proposed model architecture flow for both BERT and LLaMA 3.

4 Expected Results

We reasonably expect a performance improvement with fine-tuned LLaMA 3 achieving 85-90% accuracy compared to BERT's 80-85%, due to LLaMA 3's larger scale and advanced pre-training. Insights hoped for include interpretability via attention visualizations showing how LLaMA 3 better captures ironic patterns, robustness to noisy inputs, and fairness analysis on potential biases in sarcasm detection across headline styles.

5 Timeline

- **Week 1 (Nov 28 - Dec 5):** Data exploration, preprocessing pipeline, and establishing the BERT baseline.
- **Week 2 (Dec 6 - Dec 12):** Implementation of Llama 3 fine-tuning with LoRA and hyperparameter tuning.
- **Week 3 (Dec 13 - Dec 18):** Error analysis, comparison of confusion matrices, and final report preparation.