Fine Tuning Large Language Models for Enhancing Binary Text Classification

Siri Chandana Errabelli || Vamshi Naik Vislavath || Sai Aravind Yanamadala
se2596 vv2289 sy3902

Problem Statement:

Text classification is a crucial task in natural language processing (NLP) with wide-ranging applications such as sentiment analysis, spam detection, and topic categorization. While pre-trained language models like RoBERTa and XLNet have demonstrated exceptional capabilities in understanding language context, they aren't particularly effective at specialized binary classification tasks such as detecting malicious prompts related to prompt injection without additional training. This project aims to assess the performance of these pre-trained models on a binary text classification task designed to identify malicious prompts, highlighting their initial limitations. We will then fine-tune the models to enhance their accuracy and robustness in this specific context.

Dataset Description:

We will utilize a labeled dataset from the Hugging Face library, consisting of text prompts classified as either benign or malicious . These datasets include diverse prompt formats and language variations, ensuring comprehensive coverage.

Model Description:

We will first evaluate the base version of RoBERTa on its ability to classify prompts as malicious or benign. This initial testing will help us understand their baseline performance and identify areas of improvement. Following this, we will fine-tune the better-performing model to adapt it for binary classification. The fine-tuning process will include optimizing hyperparameters such as learning rate, batch size, and training epochs to achieve robust performance.

Goals and Target Metrics: The goal of the project is to develop a high-performance classifier for prompt vulnerabilities. We aim to achieve

1. Significant performance improvement between the pre-trained and fine-tuned models, highlighting the impact of task-specific training.
2. (Accuracy >= 90% Precision: >= 90% Recall: >= 90% F1 Score: >= 90%)

The deliverables will include both base model and fine-tuned model, a comprehensive evaluation of their performance, and recommendations for deploying the classifier in practical settings.

Literature survey:

- **Wallace, E., et al. (2019).** "Universal Adversarial Triggers for Attacking and Analyzing NLP." *EMNLP-IJCNLP.*
- ***Perez, F., & Ribeiro, I****. "Ignore Previous Prompt: Attack Techniques for Language Models." arXiv, 2022.*
- **Solaiman, I., et al. (2019).** "Release Strategies and the Social Impacts of Language Models." *arXiv preprint arXiv:1908.09203.*