Explainable Sentiment Analysis Using Prompt Engineering

Jayanth
017461483

Abstract

This project investigates how different prompting strategies—Direct, Few-Shot, and Chain-of-Thought—impact the performance of a pre-trained transformer model on the IMDB movie review sentiment classification task. Beyond predicting sentiment, the project applies Explainable AI techniques including attention visualization and Integrated Gradients to highlight which tokens influence the model's decisions. A Gradio interface is built to allow real-time testing, visualization, and comparison of prompt styles. Results show that even without retraining, simple prompt formatting can affect performance by more than 10%, and that token-level attribution improves transparency and interpretability of model behavior.

Introduction

With the rise of large language models (LLMs), prompt engineering has become a compelling alternative to fine-tuning. Rather than modifying model parameters, prompt engineering focuses on rephrasing inputs to elicit better performance. This approach is especially relevant for resource-constrained settings and tasks like sentiment analysis. In this project, we explore how varying prompt structure can impact a pre-trained model's sentiment predictions. Moreover, we incorporate Explainable AI (XAI) tools to illuminate how the model arrives at its decisions—allowing users to not only receive a sentiment label, but to understand the reasoning behind it.

Related Work

Recent literature has shown that prompting techniques like Few-Shot Learning and Chain-of-Thought (CoT) prompting can improve accuracy in classification, reasoning, and generation tasks. Works like "Language Models are Few-Shot Learners" and "Chain-of-Thought Prompting Elicits Reasoning" demonstrate how LLMs can be steered effectively with just a few examples or structured reasoning prompts. In parallel, the field of Explainable AI has developed tools like attention maps and attribution methods (e.g., Integrated Gradients) to help interpret black-box model predictions. This project combines both perspectives by evaluating prompt strategies and visualizing model behavior.

Dataset

We use the IMDB movie reviews dataset available through TensorFlow Datasets. It consists of 50,000 movie reviews labeled as positive or negative. For this project:

- We select 100 positive and 100 negative reviews from the test set (balanced subset).
- Reviews are used without modification except for truncation to fit input limits.
- These are then embedded into different prompt templates and fed to the model.

Methods

Prompting Strategies

We evaluate 3 prompt formats:

Direct

```
Review:
{text}
Sentiment?
```

Few-Shot
Three labeled examples are shown before the test review:

```
Review:
Example 1...
Label: Positive
...
Review:
{text}
Label:
```

Chain-of-Thought
Model is guided to think before answering:

```
Review:
{text}
Let's think step by step about whether this is Positive or Negative.
```

Model Used

- Hugging Face model: `nlptown/bert-base-multilingual-uncased-sentiment`
- Output: 1 to 5 stars (we map as follows):
  - 1–2 stars → Negative
  - 3 stars → Ambiguous / mapped based on confidence
  - 4–5 stars → Positive

Explainability Techniques

1. Attention Heatmaps
   - From Layer 1, Head 1 of the transformer
   - Visualizes which tokens attend to each other
2. Integrated Gradients (via Captum)
   - Applied on token embeddings
   - Attributes prediction to input tokens
   - Output: token-highlighted HTML with red intensity indicating importance

Experiments and Results

Accuracy by Prompt Style

On the 200-sample test set:

- Direct Prompt: ~82% accuracy
- Few-Shot Prompt: ~78–80%, improves on mixed sentiment
- Chain-of-Thought Prompt: ~75%, useful in some reasoning-heavy cases but introduces verbosity

Visual Outputs

- Probability bar charts show confidence in each of the 1–5 star classes
- Attention heatmaps reveal focus patterns of the model
- Integrated Gradients attribution highlights specific tokens that influenced the decision (e.g., "predictable", "brilliant", etc.)

Conclusion

Prompting strategies can dramatically affect the output of even pre-trained models without fine-tuning. Direct prompts work well for short, clear reviews. Few-shot prompting helps on ambiguous or sarcastic samples. Chain-of-Thought prompts offer potential but may confuse non-reasoning tasks. Explainability through attention and attribution gives us insight into which words matter most to the model, increasing transparency and user trust. This project demonstrates a lightweight, interpretable approach to sentiment classification using prompt engineering.