Assignment 4.2: Prompt Tuning Experiment for Sentiment Analysis

June 26, 2025

1 Introduction

This report details a prompt tuning experiment for sentiment analysis using the Flan-T5 model (`google/flan-t5-base`). The task involves classifying the sentiment of a movie review, "The movie was surprisingly good and kept me engaged till the end," into one of three classes: `positive`, `neutral`, or `negative`. The expected label is `positive`. We compare three prompt styles—Direct, Contextual, and Pattern-based—to evaluate their effectiveness in terms of Exact Match (EM) and F1 score.

2 Methodology

2.1 Task

The task requires classifying the sentiment of the input sentence: "The movie was surprisingly good and kept me engaged till the end." The ground truth label is `positive`. The model's predictions were evaluated for accuracy and generalizability across prompt styles.

2.2 Model

We used `google/flan-t5-base`, a text-to-text transformer model fine-tuned for instruction-based tasks, with its fast tokenizer. The experiment was run on a CPU (or CUDA if available) using Python 3.11, `transformers==4.45.1`, `torch==2.3.0`, and `datasets==2.21.0`.

2.3 Prompts

Three prompt styles were tested:

- **Direct Prompt**: `Sentiment of the following sentence: The movie was surprisingly good and kept me engaged till the end.`
- **Contextual Prompt**: Provides explicit options: `Analyze sentiment. Sentence: 'The movie was surprisingly good and kept me engaged till the end.' Options: positive, neutral, negative. Answer:`
- **Pattern-based Prompt**: Uses a structured format: `Text: 'The movie was surprisingly good and kept me engaged till the end.' Sentiment classification (positive / neutral / negative):`

2.4 Evaluation

Predictions were generated using Flan-T5's generative approach, with a maximum of 10 new tokens. Outputs were decoded, skipping special tokens, and evaluated using:

- **Exact Match (EM)**: 1 if the prediction matches the ground truth (`positive`) exactly, 0 otherwise.
- **F1 Score**: Harmonic mean of precision and recall based on token overlap, treating the label as a single token.

Normalization (lowercase, remove punctuation, normalize spaces) was applied for consistency.

3 Results

Table 1 summarizes the predicted answers and evaluation metrics for each prompt.

Table 1: Predicted Answers and Evaluation Metrics

Prompt Type	Predicted Answer	EM	F1
Direct	positive	1.0	1.0
Contextual	positive	1.0	1.0
Pattern-based	positive	1.0	1.0

4 Analysis

All three prompt styles correctly predicted the sentiment as `positive`, achieving perfect EM (1.0) and F1 (1.0) scores for this single test case. The Direct Prompt's simplicity allowed Flan-T5 to directly interpret the task and produce the correct label, leveraging its instruction-tuned capabilities. The Contextual Prompt, by explicitly listing options (`positive`, `neutral`, `negative`), provided additional clarity, ensuring the model focused on the classification task. The Pattern-based Prompt's structured format similarly guided the model to produce a concise, correct output. While all prompts performed equally well for this straightforward positive review, the Contextual and Pattern-based prompts are expected to generalize better for more ambiguous or complex reviews, as they provide explicit task framing. No errors were observed in this case, but challenges may arise with neutral or negative sentiments, where nuanced language could lead to misclassifications.

5 Conclusion

The experiment demonstrates that all three prompt styles—Direct, Contextual, and Pattern-based—effectively classified the sentiment of the given movie review as `positive`, achieving perfect EM and F1 scores. For simple sentiment analysis tasks, the Direct Prompt is sufficient due to its minimal structure and alignment with Flan-T5's capabilities. However, the Contextual and Pattern-based prompts are likely to offer better generalizability for diverse inputs, as their structured formats enhance task clarity. Future work could evaluate these prompts on a larger dataset with varied sentiments to assess their robustness and scalability.