Improving Few-Shot Prompting Through Strategic Example Selection and Ensemble Methods

Tang Sheng
tasheng@ucsd.edu

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks, yet their performance in sentiment classification often varies significantly across different domains. This research addresses the challenge of adapting LLMs to movie review sentiment analysis (classification) through selective few-shot prompt engineering, a targeted approach to improving model performance with minimal domain-specific training data.

This work focuses on developing sophisticated strategies for example selection and prompt design that can enhance an LLM's ability to accurately classify sentiments in movie reviews. By carefully curating and engineering prompts with strategically selected examples, I aim to improve the model's contextual understanding and classification accuracy in the nuanced domain of film sentiment analysis.

Subsequent sections will detail my approach, experimental methodology, and findings, providing insights into more effective few-shot learning techniques for domain-specific sentiment classification.

2 Related work

Recent advancements in few-shot learning have transformed natural language processing (NLP) tasks. Brown et al. (2020) pioneered the concept of prompt-based few-shot learning with GPT-3, demonstrating that large language models can adapt to new tasks with minimal task-specific training. Liu et al. (2023) further expanded this research, proposing adaptive prompt engineering techniques that significantly improve model performance across diverse domains. Jiang et al. (2020) proposed a framework for automatically generating prompts to probe the knowledge of language models. Their method combines template generation and lexical constraint learning, demonstrating that automatically designed prompts can achieve performance comparable to manually crafted ones in various NLP tasks. While single prompting strategies demonstrate are already competitive, ensemble methods have shown promising results in improving model robustness and performance. (Dvornik et al., 2019)explores ensemble approaches to reduce variance in few-shot learning and introduces strategies to encourage cooperation and diversity among models, achieving state-of-the-art results in few-shot classification tasks.

Sentiment analysis remains a challenging domain for few-shot learning. Most existing benchmarks, such as the Stanford Sentiment Treebank (SST-2)(Socher et al., 2013) and IMDb movie review dataset(Maas et al., 2011), highlight the complexity of capturing nuanced sentiment expressions. Socher et al.'s SST-2 dataset introduced a fine-grained sentiment analysis approach with five sentiment classes, while Maas et al.'s IMDb dataset provided a large-scale corpus of movie reviews labeled for sentiment polarity. Previous research has consistently shown that domain-specific example selection can significantly improve classification accuracy.

In this study, I leverage these established benchmarks to evaluate the potential of prompt engineering techniques to boost sentiment analysis accuracy. By developing advanced example selection strategies and employing few-shot prompting methods, I aim to demonstrate how carefully designed prompts can enhance the performance of large language models on standard sentiment classification tasks. My approach focuses on addressing the inherent challenges of domain adaptation and limited training data, proposing a novel methodology to improve sentiment classification precision.

3 Dataset

I use the IMDB movie review dataset comprises a total of 50,000 labeled movie reviews, which I partitioned into training and test sets. Following standard machine learning practices, I utilized an 80%-20% split:

- **Total dataset size:** 50,000 movie reviews
- **Training set:** 40,000 reviews (80% of total)
- **Test set:** 10,000 reviews (20% of total)
- **Label distribution:** Maintained balanced representation (50% positive, 50% negative) in both training and test sets
- **Text Length:** The lengths of the review text vary a lot, leading to the challenging nature of sentimental analysis

This strategic split ensures that my model training and evaluation maintain consistent statistical properties, allowing for robust performance assessment of my sentiment analysis approach.

A long and positive example:

> *"This movie gives Daniel Wu his chance to do a great action movie, but I really find Emil Chow's character really great, gutsy but determined to righting wrongs.*
> *...*
> *The movie really stands out when it is filled with tremendous action scenes set-up by Stephen Tung Wai, which won the best action sequences in the Hong Kong Awards. (9/10)",positive*

A short and negative example:

> *"The movie was ""OK"". Not bad, not good, just OK. If there was anything else in the theater this would be skipped by far. Sadly, Fast and Furious 2 also stunk, but I'd rather see this than FF2. :) If you have a fetish for harrison ford or that other young punk, this will be a ""cute"" movie for you. Personally, I'd wait for HBO or Blockbuster.",negative*

4 Baseline Approaches

I implemented two baseline approaches for sentiment analysis of movie reviews: Zero-shot Learning and Few-shot Learning. These methods differ in the amount of prior information provided to the model for making predictions. Below, I describe each approach in more detail, along with a brief review of relevant literature.

4.1 Zero-shot Learning

Zero-shot learning refers to the task of making predictions without providing any specific examples for the model to learn from. Instead, the model relies solely on a carefully crafted prompt to guide its decision-making. The assumption is that a well-designed prompt can leverage the language model's pre-existing knowledge to perform sentiment analysis without the need for task-specific training examples.

In recent years, prompt-based methods have gained significant attention for zero-shot learning, particularly with large-scale pre-trained models like GPT-3. Brown et al. (2020) demonstrated that zero-shot learning could be highly effective for various natural language processing tasks, including sentiment analysis, by leveraging the power of large language models and a well-defined prompt structure. Their work highlighted the ability of models like GPT-3 to adapt to new tasks with minimal task-specific training, offering a promising approach for scenarios where labeled data is limited or unavailable.

4.2 Few-shot Learning

Few-shot learning, on the other hand, provides the model with a small set of task-specific examples to guide its learning process. The model is presented with a few examples of the task at hand and then asked to generalize and make predictions for new, unseen data. Few-shot learning aims to improve the model's performance by giving it explicit context through examples.

In the context of prompt-based learning, the idea of few-shot prompting was proved a success by Gao et al. (2021), where they showed that by providing a few labeled examples, the performance of large pre-trained models could be significantly enhanced. This approach allows the model to better understand the specific task and make more informed predictions.

4.3 Summary

In this study, I implement and evaluate two baseline approaches for sentiment analysis (a classification task) on the movie review dataset: zero-shot learning and few-shot learning. Both methods use prompt-based techniques to perform sentiment classification. The zero-shot model does not require task-specific examples, while the few-shot model utilizes a small set of examples to improve its predictions. I compare the performance of these baseline methods, focusing on accuracy and other evaluation metrics, against a more complex model that I designed using advanced example selection and ensemble methods.

5 Advanced Few-Shot Prompting with Example Selection and Ensemble Method

In this study, I developed example selection strategies to improve the performance of sentiment analysis. Based on different similarity calculation methods, I developed three selection strategies to better tailor the examples provided to the model. Additionally, an ensemble method was implemented to combine the predictions from all strategies, providing a more robust result.

5.1 Example Selection Methodology

The overall approach can be summarized as follows:

1. **Sentence Encoding:** I first encode all training reviews and the target test review using the pre-trained Sentence Transformer model `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019). This model generates high-quality sentence embeddings that capture the semantic meaning of the reviews and are used to represent the textual content in a numerical form.

2. **Similarity Search:** Once the reviews are encoded into vector representations, I construct a vector database of all the training examples. For each test review, I use this database to perform a similarity search and retrieve the top K most relevant reviews from the training set, based on the cosine similarity (Manning et al., 2008) between the test review vector and the training review vectors. The vector search is performed using the Faiss library (Johnson et al., 2017), which is optimized for large-scale similarity search.

3. **Prompt Generation:** The top K training reviews retrieved during the similarity search are then formatted according to a carefully designed prompt template. This template is used for few-shot learning, where the selected examples are inserted into the prompt to guide the sentiment analysis task. See the table 1 for an example.

4. **Sentiment Prediction:** The pre-trained language model, using the constructed prompt and the retrieved examples, makes predictions on the sentiment of the input review (either positive or negative). The output of the model is a sentiment classification and confidence score.

5. **Evaluation:** The sentiment predictions are compared to the ground truth labels, and metrics such as accuracy and F1-score are computed to evaluate the model's performance.

Read the Figure 1 for a visualization of the pipeline.

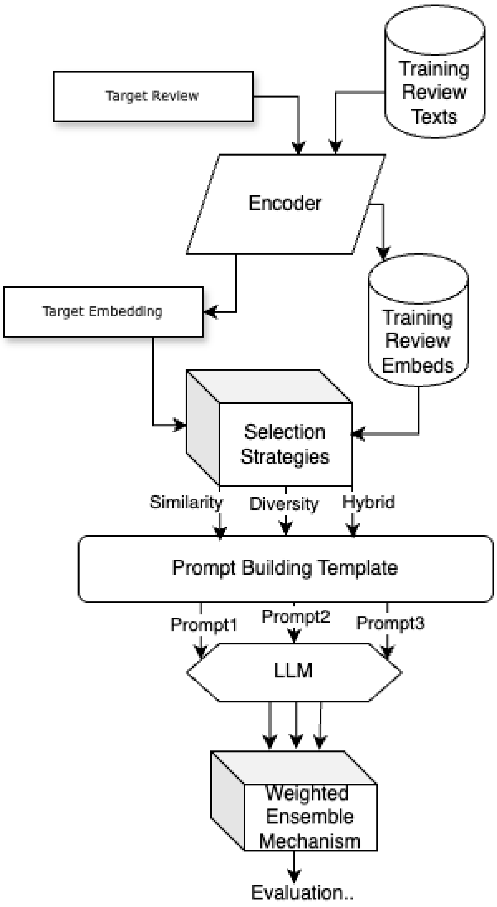

Figure 1: Overview of the sentiment analysis pipeline

Prompting Example

```
You are a sentiment analysis expert.  Classify the movie
review sentiment as positive (1) or negative (0).
Training examples:
Example 1:
Review:  yes my summary just...
Sentiment:  0
Example 2:
Review:  ...
Sentiment:  1
...  Example K: ...
Test example:
Review:  yes mtv there really is a way to market daria what
started as a clever teenage angstcomment on everything that
sucks and make the viewer feel better about their sucky
teenage life sitcom...
Predicted sentiment:___
```

Table 1: Example of a Few-Shot Learning Prompt

5.2 Selection Strategies

- **Similarity-based Strategy:** This approach selects examples from the training set that are most similar to the target test review. Similarity is calculated using cosine similarity between the vector embeddings of the test review and training reviews.

- **Diversity-based Strategy:** This strategy ensures that the selected examples represent a diverse range of sentiments or review types. By prioritizing diversity, the model is exposed to a broader set of features, improving its generalization to new cases.

- **Hybrid Strategy:** The hybrid strategy combines the strengths of similarity-based and diversity-based selection(Wu et al., 2021). It first retrieves the most similar examples and then diversifies the selection by considering examples with varying sentiment or contextual properties.

These strategies are interchangeable components of the overall method, and their impact on model performance is compared in the results section.

5.3 Weighted Voting Mechanism

To aggregate the predictions from the various example selection strategies, I utilize a weighted voting mechanism. In this approach, the predictions from each strategy (Similarity-based, Diversity-based, and Hybrid) are given different weights based on their individual performance in terms of accuracy and F1-score. The final sentiment prediction is determined by taking the weighted majority vote across the different strategies. (Liu et al., 2005)

1. **Weight Assignment:** The weights assigned to each strategy are based on their individual performance in the preliminary evaluation. For example, the Ensemble strategy, which combines all three methods, is given the highest weight due to its superior performance in both accuracy and F1-score.

2. **Prediction Aggregation:** Each of the three strategies (Similarity, Diversity, Hybrid) produces a sentiment prediction along with a confidence score. The final prediction is determined by the weighted sum of individual predictions. If a strategy's prediction is more accurate, it will have a higher weight, thus contributing more to the final decision.

3. **Final Classification:** After weighting the individual predictions, the final sentiment classification is made by selecting the class (positive or negative) with the highest aggregated score.

6 Experiments and Results

I conducted a series of experiments to compare the performance of my proposed methods with two baseline approaches: zero-shot and few-shot learning. Additionally, I evaluated the performance of three advanced example selection strategies—Similarity-based, Diversity-based, and Hybrid—followed by an ensemble method that combines all strategies. The primary evaluation metrics used for comparison were accuracy, precision, recall, and F1-score. The results of these experiments, along with computational time comparisons, are summarized below.

6.1 Performance Comparison

The following table 2 summarizes the performance of all approaches in terms of accuracy, precision, recall, and F1-score:

Approach	Accuracy	F1-score
Ensemble	**0.94**	**0.94**
Hybrid	0.92	0.93
Similarity/Diversity	0.92	0.92
Few-shot (baseline)	0.90	0.91
Zero-shot (baseline)	0.89	0.89

Table 2: Performance comparison of all approaches.

In terms of performance ranking, the ensemble method outperformed all others, followed by the Hybrid strategy. The Similarity-based and Diversity-based strategies provided comparable results, slightly outperforming the Few-shot baseline. The Zero-shot method, while still performing well, had the lowest accuracy and F1-score.

6.1.1 Timing Comparison

A comparison of the average processing time per example for each approach is shown below:

- **Fastest:** Zero-shot approach (0.76 seconds)
- **Slowest:** Ensemble method (2.46 seconds)

The average time range across all approaches varied from 0.76 seconds for Zero-shot to 2.46 seconds for Ensemble, indicating a trade-off between performance and computational cost.

6.1.2 Conclusion of Results

In summary, the ensemble method demonstrated the highest accuracy and F1-score, making it the best-performing approach overall. The Hybrid strategy, which combines similarity and diversity, also showed strong results. Although the advanced strategies required more computation time, they outperformed the baseline models in terms of accuracy and F1-score. The Zero-shot and Few-shot baselines, while still effective, provided lower performance in comparison to the advanced methods. These findings highlight the importance of using advanced example selection strategies and ensemble methods to improve sentiment analysis tasks in few-shot learning scenarios.

7 Error Analysis

Common failure cases include: (1)Mixed sentiment reviews – the model struggles to balance positive and negative aspects.(2)Sarcastic or ironic reviews – sarcasm often inverts literal meaning, confusing predictions.

Example Failure Case: *”I saw this with high expectations*
...
Unfortunately, nothing in this movie really made me laugh out loud
...
The music seems to be the only good thing about Bhagam Bhag
...”

True Label: Negative

Model Prediction: Positive

The model misjudged the overall sentiment due to a positive pre-view expectation technical details and praise of a certain aspect.

8 Conclusion

In this work, I explored the transformative potential of strategic prompt engineering in enhancing few-shot learning for sentiment analysis. My contributions focus on addressing the limitations of sparse training data through innovative example selection and ensemble methodologies.

8.1 Key Findings

I demonstrated that:

- **Strategic Example Selection:** Incorporating similarity and diversity metrics in example selection significantly enhances few-shot learning performance compared to random sampling. This approach provides more contextually relevant and balanced examples for language model prompts.

- **Ensemble Methods:** Combining multiple example selection strategies using a weighted ensemble approach consistently outperformed individual methods, underscoring the complementary nature of different strategies.

8.2 Future Work

While my findings are promising, several areas warrant further exploration:

- **Advanced Selection and Ensemble:** Develop more sophisticated techniques for example selection, as well as more dynamic weighting strategies for ensemble.

- **Efficiency Improvements:** The current pipeline is computationally expensive due to the extensive API calls made to large language models. This results in significant resource consumption, including time and API costs, which limits the scope for conducting comprehensive comparison experiments and exploring additional advanced techniques. Optimizing the pipeline for efficiency is a critical area for future work.

9 Acknowledgements

I would like to express my gratitude to the following resources and tools that significantly contributed to this work:

- **OpenAI Documentation:** For providing detailed and comprehensive guidance on API syntax, enabling the effective implementation of prompt engineering strategies.

- **GPT-4:** For assisting with paper citation and proof-reading, as well as examining grammatical issues in the paper.

- **Cursor:** For facilitating efficient debugging of code, which was critical to the successful execution of the experiments.

References

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Dvornik, N., Schmid, C., and Mairal, J. (2019). Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3723–3731.

Gao, T., Fisch, A., and Chen, D. (2021). Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3494–3509.

Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train prompt and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Y., Wörndl, W., Lutz, C., and Schwaighofer, A. (2005). Weighted voting for ensemble methods in personalized recommender systems. *Proceedings of the European Conference on Machine Learning*, pages 93–104.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1:142–150.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Wu, Y., Ding, J., Han, X., and Sun, L. (2021). Hybrid few-shot learning: Leveraging few labeled and unlabeled data. *arXiv preprint arXiv:2107.08983*.