Data Annotation and Sentiment Analysis of Movie Reviews Using Zero-Shot Classification

Sriram Theerdh Manikyala
manikyala@arizona.edu

Abstract

In this project, I assess the results of using the facebook/bart-large-mnli zero-shot classification model to perform sentiment analysis on a dataset of IMDb movie reviews that I constructed and annotated with a fellow annotator. The gold labels were obtained by merging annotations from both annotators, and this dataset has a Cohen's Kappa Score of 0.912, indicating good agreement. The dataset was pre-processed through systematic cleaning. Domain-specific fine-tuning was avoided to evaluate the model's performance. The classification was performed using the cleaned data, and the model's performance was assessed with metrics like accuracy and F1 score. The results show that the model performs well, highlighting the importance of collaborative annotation and systematic preprocessing for machine learning tasks.

1 Introduction

Sentiment analysis is a crucial task for understanding sentiment in reviews, social media, and other types of textual data. This project aims to classify IMDb movie reviews into two sentiment categories: Positive/Neutral and Negative. The project also emphasizes the importance of sentiment analysis in various fields like entertainment and marketing. Even though models pretrained on larger corpora, such as facebook/bart-large-mnli, show strong performance in zero-shot tasks, the performance is influenced by the quality of the input data. This study examines the effects of systematic data cleaning and collaborative gold label generation on model performance.

2 Methodology

2.1 Data Annotation

I collected a dataset of 250 IMDb movie reviews, and the first step was annotating the data. We classified the reviews as:

- '0': Positive/Neutral – Reviews that are very positive, express mixed views leaving mostly positive comments, or are somewhat negative but end on a supportive note.
- '1': Negative – Reviews that contain

fierce criticism, dissatisfaction, or emotional disappointment.

I was Annotator-1, and my classmate, Sai Hemanth Kilaru, was Annotator-2. We both followed a well-defined annotation guideline to maintain consistency in labeling. The gold labels were derived from the annotations of both annotators, producing a reliable final dataset for the task.

Table 1: Distribution of Annotations

Label	0	1
Annotator 1	125	125
Annotator 2	126	124
Gold Label	128	122

2.2 Evaluation of Annotator Agreement (Cohen's Kappa Score)

Cohen's Kappa Score was calculated to measure the reliability of the annotation using the `cohen_kappa_score` metric from scikit-learn. This metric evaluates how much more agreement between Annotator-1 and Annotator-2 there is than would be expected by chance. A score of 0.912 was achieved, indicating nearly perfect agreement between the two annotators. This confirmed that the annotations were consistent and reliable, allowing us to use them as gold labels for the final classification.

2.3 Data Pre-Processing

The dataset was pre-processed to improve its quality before modeling. The cleaning steps included:

- Replacing NaN reviews with empty strings.
- Converting text to lowercase.
- Removing white spaces at the beginning and end.
- Removing special characters and punctuation.
- Normalizing whitespace.

These steps standardized the text without removing valuable information, and the cleaned dataset was saved as `cleaned_dataset.xlsx` for further analysis.

2.4 Setting Up the Model and Performing Zero-Shot Classification

We used Hugging Face's `facebook/bart-large-mnli` model to evaluate the cleaned dataset. We selected this pre-trained model because it is suitable for zero-shot classification tasks. The model was applied directly, without fine-tuning, to predict sentiment labels for individual reviews. Candidate labels: "Positive" and "Negative." If the model predicted a higher probability for "Positive," the review was labeled 0. If "Negative" had a higher probability, it was labeled 1. The predicted labels were stored alongside the gold labels for evaluation.

2.5 Model Evaluation

Once the model generated its predicted labels, they were compared to the gold labels

defined by Annotator-1 and Annotator-2. Performance evaluation metrics like accuracy and F1 score were computed to assess how well the model matched the human annotations. These metrics confirmed that the model was able to accurately classify reviews on a clean dataset.

3 Evaluation Metrics

Accuracy: The ratio of correctly predicted labels, indicating how often the model predicted correctly compared to the ground truth.

F1 Score: The harmonic mean of precision and recall, offering a balance between false positives and false negatives. This is crucial in binary classification tasks like this one, where class imbalance may influence performance.

Table 2: Model Performance Metrics

Metric	Score
Accuracy	94.4%
F1 Score	0.944

4 Conclusion

This project demonstrates the effectiveness of using a pre-trained model, `BART-large-MNLI`, for sentiment classification in a zero-shot setting. The model achieved high accuracy and F1 scores without the need for domain-specific fine-tuning. The results emphasize the importance of clean, well-annotated data in boosting model performance. The combination of annotation consistency, data preprocessing, and a powerful pre-trained model contributed to the successful sentiment classification of IMDb movie reviews.

References

1. Reviews taken from `https://www.imdb.com/?ref_=nv_home`
2. TensorFlow. (2023). Text Classification with Movie Reviews. TensorFlow Hub. Retrieved from `https://www.tensorflow.org/hub/tutorials/tf2_text_classification`
3. Dataiku. (2023). GPT-based zero-shot text classification with the OpenAI API. Dataiku Developer Documentation. Retrieved from `https://developer.dataiku.com/12/tutorials/machine-learning/genai/nlp/gpt-zero-shot-clf/index.html`