Sentiment Analysis on Restaurant Reviews

Shivank Singh Thakur, Rahul Reddy Gangapuram, Prudhvi Sai Raj Dasari, Jathin Shettigar Nagabhushan
Department of Computer Science,
San Jose State University
{shivanksingh.thakur, rahulreddy.gangapuram, prudhvisairaj.dasari, jathin.shettigarnagabhushan}@sjsu.edu

***Abstract*—In this project, we perform the task of sentiment analysis on restaurant reviews using NLP techniques, pre-trained models and classic machine learning techniques. The goal is to classify reviews into binary classes: positive and negative sentiments. The dataset is split into 80% training and 20% testing subsets with balanced sentiment distributions. We experiment with three different models, including deep learning-based pre-trained models and traditional classifiers enhanced with text embeddings. The three models used are Logistic regression and Random Forest with TF-IDF word embeddings, and fine-tuned BERT model. This paper discusses hyperparameter tuning, model fine-tuning, and evaluation metrics (accuracy and F1-score), and presents a comparative analysis of the performance of the three models. Our results show that fine-tuned BERT outperforms the other models, achieving 96% accuracy and 0.96 F1 score on the test set.**

***Index Terms*—Sentiment Analysis, Restaurant Reviews, BERT, TF-IDF, Random Forest, Text Embeddings.**

I. Introduction

Online restaurant reviews play a critical role in shaping consumer decisions, making it essential to understand the sentiment behind each review. In this project, we perform sentiment analysis by classifying reviews into positive and negative sentiments using pre-trained models and classic machine learning approaches. This paper explores different model architectures and evaluates their effectiveness in terms of accuracy and F1 score. [1]

A. Key Contributions

- Rahul and Jathin worked on data preprocessing, exploratory data analysis and word embeddings.
- Shivank and Prudhvi worked on model training, fine-tuning and testing.
- We used feature engineering techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) and pre-trained embeddings like BERT, to extract meaningful features from the restaurant reviews. We explored different classic and pre-trained models and compared them to evaluate their effectiveness in classifying the restaurant review sentiment.

B. Organization of the Paper

The paper is organized as follows: Section II describes the dataset used, Section III outlines the preprocessing steps, Section IV explains the models used, Section V presents the experimental results, and Section VI concludes the paper with a discussion of future work.

[1]https://www.overleaf.com/read/mjcbjzgzzzvf#64c375

II. Dataset Description

The dataset consists of restaurant reviews, each labeled as either positive (1) or negative (0). We used the balanced dataset to ensure equal representation of positive and negative reviews. The data was split into an 80% training set and a 20% testing set to evaluate model performance.

III. Preprocessing and Embedding

We performed the following preprocessing steps for the sentiment analysis task:

1) **Data Shuffling and Splitting:** The dataset was shuffled and split into 80% training and 20% testing subsets while maintaining the sentiment distribution. Additionally, the dataset is stratified, which ensures that both the training and testing sets have approximately the same proportion of positive and negative reviews as the original dataset.
2) **Preprocessing:** Depending on the chosen model, we performed necessary preprocessing. Some models required raw text, and others utilized text embeddings. The reviews were preprocessed using the TfidfVectorizer from scikit-learn, which converts the text data into TF-IDF feature vectors. We used a maximum of 5000 features for the vectorization process.

Fig. 1: Word Cloud for Positive Reviews

The restaurant reviews dataset likely consists of relatively clean, user-generated text without much noise or complex structures. The use of TF-IDF vectorization inherently handles some basic preprocessing tasks, such as tokenization and conversion to lowercase. The BERT model has its own tokenizer and preprocessing pipeline, which is designed to handle raw text input effectively.

Fig. 2: Word Cloud for Negative Reviews

While more extensive preprocessing (such as removing stopwords, stemming, or lemmatization) was performed, however, we didn't see any increase in the model performance.

IV. Classification Models

We employed three models for sentiment analysis:

- **Logistic Regression with TF-IDF:** A logistic regression model was trained using TF-IDF representations of the reviews to predict sentiment. This approach captures the importance of words in reviews without requiring deep learning models. The model was optimized using RandomizedSearchCV with hyperparameters including regularization strength (C), maximum iterations, penalty type, and solver.
- **Random Forest with TF-IDF:**A random forest classifier was trained using TF-IDF representations of the reviews. This ensemble approach combines multiple decision trees for improved robustness. The model was optimized using RandomizedSearchCV with hyperparameters including maximum depth, maximum features, minimum samples leaf, minimum samples split, and the number of estimators.
- **Pre-trained BERT Model:** A BERT (Bidirectional Encoder Representations from Transformers) model was fine-tuned for the sentiment analysis task. BERT is a deep learning-based model that captures contextual information in text.

A. Evaluation Metrics

We evaluated the models based on two key metrics:

- **Accuracy:** The ratio of correctly predicted sentiment labels (positive or negative) over the total number of predictions. This metric provides an overall measure of the model's performance across both classes.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of the model's performance on both classes. This metric is particularly useful for imbalanced datasets, as it takes into account both false positives and false negatives.
- **Confusion matrix:** This matrix allows us to identify specific areas where each model excels or struggles, providing insights into potential biases or weaknesses.
  - True Positives (TP): Correctly identified positive reviews
  - True Negatives (TN): Correctly identified negative reviews
  - False Positives (FP): Negative reviews incorrectly classified as positive
  - False Negatives (FN): Positive reviews incorrectly classified as negative

V. Results and Analysis

The following table summarizes the performance of the models on the train and test set:

Model	Train Accuracy	Train F1-score	Test Accuracy	Test F1-score
Logistic Regression	0.9904	0.9908	0.9423	0.9434
Random Forest	0.9712	0.9722	0.8462	0.8519
BERT	0.9952	0.9953	**0.9615**	**0.9615**

TABLE I: Model Performance

Figures 3, 4 and 5 show the confusion matrices for the Logistic Regression, RandomForest and Pre-trained BERT models, respectively. The BERT model performed well due to its ability to understand contextual relationships in text, achieving the highest F1-score.

In this sentiment analysis project on restaurant reviews, we trained and compared three different models: Logistic Regression, Random Forest, and BERT. The results demonstrate that all three models performed well, with BERT achieving the highest accuracy and F1 score on both training and test sets.

The BERT model outperformed the traditional machine learning approaches, achieving 96.15% accuracy and F1 score on the test set. This good performance can be attributed to BERT's pre-training on large text corpora and its ability to capture contextual information and nuances in language.

Logistic Regression showed strong performance as well, with 94.23% accuracy and 0.94 F1 score on the test set. This indicates that even a simpler linear model can effectively capture sentiment in restaurant reviews when combined with appropriate feature extraction techniques like TF-IDF.

The Random Forest model, while still performing reasonably well, lagged behind the other two models with 84.62% accuracy and 0.85 F1 score on the test set. This suggests that the non-linear decision boundaries created by Random Forest may not have been as effective for this particular task compared to the other approaches as it is clearly overfitting on the dataset. This overfitting might be due to the model's complexity, and it may need further exhaustive hyperparameter tuning or more diverse training data to help it generalize better to unseen examples.

VI. Conclusion and Future Work

In this work, we demonstrated the use of pre-trained BERT and classic machine learning models for sentiment analysis on restaurant reviews. The BERT model outperformed the

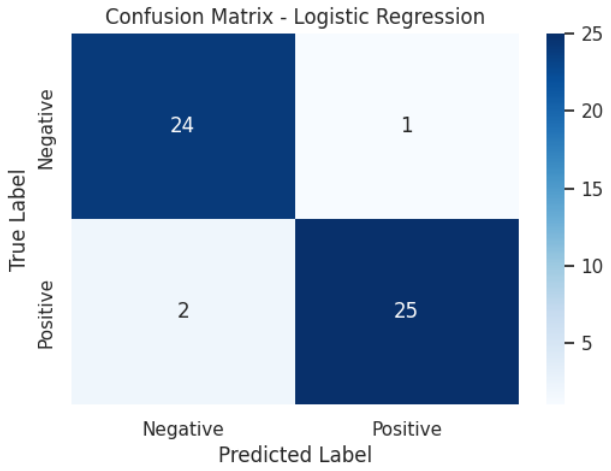

Fig. 3: Confusion Matrix for Logistic Regression + TF-IDF

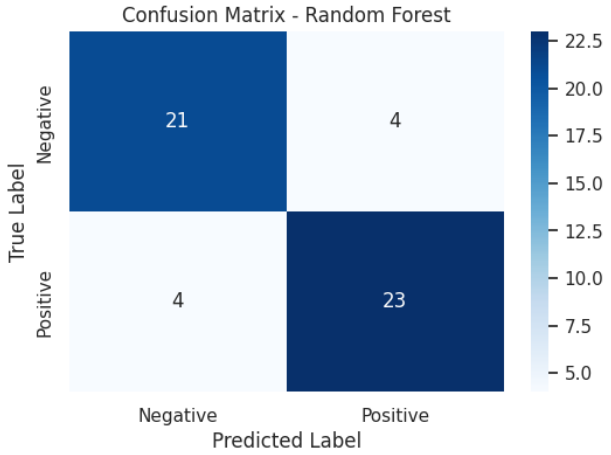

Fig. 4: Confusion Matrix for Random Forest + TF-IDF

other models, achieving the highest accuracy and F1-score, indicating its strong ability to understand textual data. However, the logistic regression also provided a good performance, showing that simple models can also fit well on data with word embeddings like TF-IDF.

Next, we plan to:

- Model Optimization: Further fine-tuning of BERT and using other variations and pre-trained models, as well as exploring advanced techniques such as transfer learning to improve performance.
- Feature Engineering: Using more sophisticated features, such as syntactic and statistical frequency or other linguistic cues, to improve accuracy.
- Hybrid Approaches: Combining deep learning with rule-based systems to better handle complex linguistic nuances.

This project shows that using deep learning models and traditional approaches can lead to highly accurate sentiment analysis systems for real-world applications.

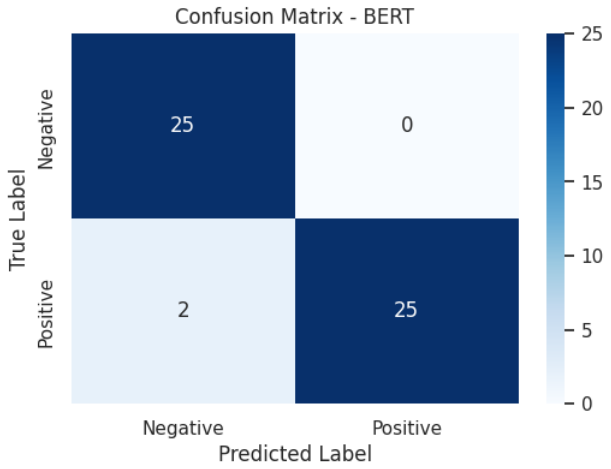

Fig. 5: Confusion Matrix for Pre-trained BERT Model

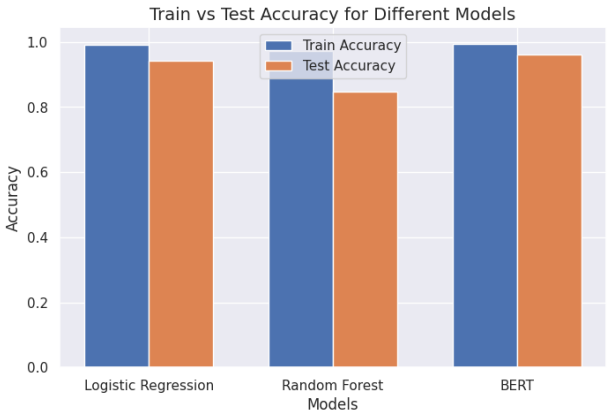

Fig. 6: Train vs Test Accuracy for Different Models