Sentiment Analysis of TripAdvisor Hotel Reviews

Lyann Sun
University of Michigan
Ann Arbor, United States
https://github.com/lyann-s/stats507FinalProject

***Abstract*—Performing sentiment analysis on hotel reviews can help companies understand the guest experience. In this project, three models- Logistic Regression, BiLSTM and a pre-trained DistilBERT model- are implemented to classify reviews from a TripAdvisor hotel dataset as negative, neutral or positive sentiment. For the Logistic Regression model, TF-IDF was used to transform the text into numerical features. GloVe was used to generate the embedding matrix for the BiLSTM model, while a pre-trained tokenizer was used to generate the contextual embeddings for the DistilBERT model. The results showed that the DistilBERT model performed best with the highest accuracy and macro F1 score, indicating that a pre-trained transformer can enhance sentiment classification tasks.**

I. Introduction

A. Background

When choosing a hotel to stay at for a trip, people often check the online reviews through different websites such as TripAdvisor, Google Reviews and Yelp. Through these reviews, guests can express their satisfaction or dissatisfaction over their stay. Different hotels may have similar ratings, but reviews can provide additional information a numerical rating cannot provide, making these reviews an important factor for people to consider when selecting a hotel.

By performing sentiment analysis on the reviews, hotel companies can understand whether customers were satisfied, unsatisfied or neutral over their stay. This information can then be used to improve the guest experience and potentially marketing strategies.

B. Project Goal

The data used in this project is a TripAdvisor hotel dataset that contains reviews and the overall rating. The goal of this project is to classify these reviews as positive, negative and neutral using different models and compare the results.

To achieve this, I implemented Logistic Regression as the baseline model and then progressed to BiLSTM and the pre-trained DistilBERT model.

C. Prior Work

Customer sentiment analysis for hotels has been studied before with different models and datasets. For example, one study analyzed 1,000 hotel reviews from Datafiniti's Business Database and compared several machine learning algorithms for hotel sentiment classification. It compared more conventional machine learning models like Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree and BernoulliNB, alongside more advanced deep learning models such as LSTM and BERT. After finishing the implementation of these and comparing the results, the study found that the deep learning models, particularly BERT and LSTM, had a better performance in comparison to the traditional machine learning methods [1].

Furthermore, there is a paper that includes a discussion and comparison on the different sentiment classification approaches applied in the hospitality industry. This review covered the approaches based on the lexicon and on machine learning, as well as hybrid strategies. Based on this review, the paper highlighted that Aspect-Based Sentiment Analysis (ABSA), which evaluates the sentiment towards a specific "aspect" of a hotel, is suitable for hotel review sentiment analysis. Moreover, it concluded that it is important to further research deep learning and unsupervised methods for sentiment analysis [2].

Based on prior work, there are multiple models and strategies that could be used to perform sentiment analysis on the hotel dataset. To evaluate their performance in sentiment analysis using the TripAdvisor dataset, in this project I implemented a traditional machine learning model, a deep learning model, and a pre-trained deep learning model, and compared the results.

II. Method

A. Problem Formulation

The task in this project is to train the Logistic Regression, BiLSTM and the DistilBERT models such that given a hotel review, it will be able to accurately classify it as one of the following classes: negative, neutral, or positive.

To tune hyperparameters, prevent overfitting and evaluate model performance, the dataset was divided into training, validation and test sets. The validation set was used to evaluate the model during training and guide the best model selection based on the macro F1 score. The chosen best model was then evaluated on the test set.

The model performance was assessed using the macro F1 score, F1 score per class, precision, recall and accuracy. Confusion matrix heatmaps were generated to provide more information about the classification errors.

B. Dataset description

The dataset used in this project can be found on https://huggingface.co/datasets/jniimi/tripadvisor-review-rating. It consists of 201,295 hotel reviews and ratings collected from TripAdvisor. The columns used in this project are “review” and “overall”, which represent the title and review, and the 1.0 to 5.0 hotel rating, respectively. Fig. 1 shows that most reviews are positive:

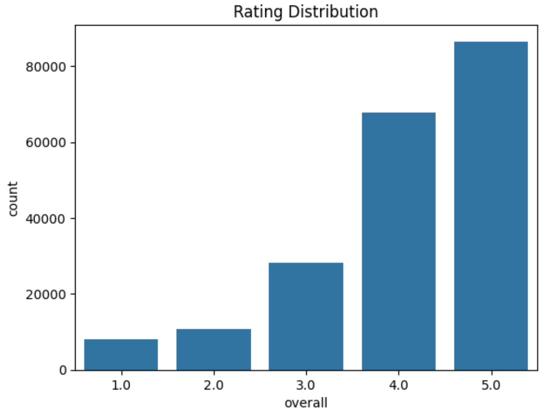

Fig. 1. Rating distribution.

To perform sentiment analysis, a new column “sentiment” was added. If the “overall” rating was 1.0 or 2.0, this was considered negative sentiment, and the corresponding value in the “sentiment” column was 0. If “overall” was 3.0, this was classified as neutral and the corresponding value under “sentiment” was 1. Lastly, if “overall” was 4.0 or 5.0, this was considered positive sentiment and the corresponding “sentiment” value was 2.

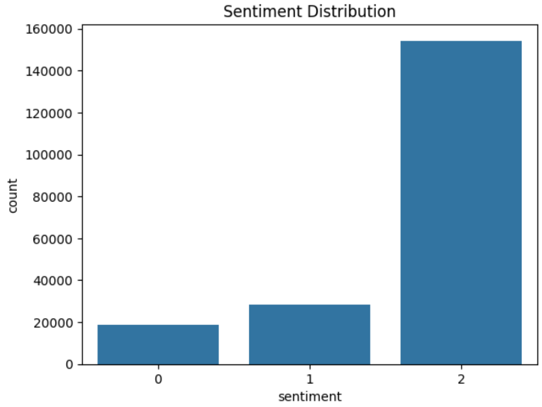

Fig. 2. Sentiment distribution.

After adding the “sentiment” column, Fig. 2 demonstrates that the sentiment distribution is heavily skewed towards the positive sentiment, meaning that the neutral and negative sentiments are underrepresented and there is a class imbalance.

C. Model Formulation

1) Logistic Regression: The Logistic Regression model required the text reviews to be transformed into numerical features. To accomplish this, TF-IDF, specifically the TfidfVectorizer tool, was utilized. When fitting the model, the maximum number of iterations allowed for the optimization algorithm to converge was set to 1,000, and the class weight was set to “balanced”. By setting the class weights to “balanced”, this adjusted the loss function to address the class imbalance. Different regularization strength values were explored during training.

2) BiLSTM: BiLSTM extends the traditional LSTM model by processing the data in both directions- forward and backward, allowing the model “to capture more contextual information” [3]. This model required the reviews to be in the form of word embedding matrices. To accomplish this, the reviews were converted to sequences, padded to a maximum length of 400, and each token was mapped to the corresponding pre-trained GloVe 300-dimensional vector to create the embedding matrix.

The BiLSTM model contained an embedding layer, a bidirectional LSTM with 128 features in the hidden state vector, a dropout layer (rate of 0.4) and a linear layer. Each review was first embedded and processed in both directions in the LSTM layer. Then, dropout was applied before the linear layer transformed the features into the final output.

An Adam optimizer was added with a learning rate of 1×10^{-3}. The criterion was set to cross entropy loss.

The model was trained for 5 epochs on GPU if available, with data loaded in batches of 32.

3) DistilBERT: DistilBERT is a “distilled version of BERT” that “reduces the size of the BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster” [4]. The pre-trained distilBERT transformer model (“distilbert-base-uncased”) was used from Hugging Face. Since this model required the text reviews to be tokenized and transformed into contextual embeddings, the DistilBertTokenizerFast tokenizer was used to preprocess the reviews.

The model was trained with the following parameters: 5 epochs on GPU if available, learning rate of 1×10^{-3}, batch size of 32, weight decay of 0.01, and evaluation at the end of each epoch. Since there were three classes, the criterion was set to cross-entropy loss by default. The best model based on the macro F1 score was saved.

III. Results

A. Logistic Regression

TABLE I
Validation Results

Inverse Regularization Parameter (C)	Macro F1-Score	Accuracy
0.001	0.64	0.77
0.01	0.69	0.80
0.1	0.73	0.83
1	0.74	0.84
10	0.73	0.84

The Logistic Regression model required 3 minutes to complete training. Different inverse regularization parameter (C) values were explored, with C = 1 yielding the highest macro F1 score after validation as observed on Table I.

TABLE II
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score
Negative (0)	0.70	0.82	0.76
Neutral (1)	0.47	0.69	0.56
Positive (2)	0.97	0.87	0.92
Accuracy			0.84
Macro Avg	0.71	0.79	0.74
Weighted Avg	0.88	0.84	0.85

Table II shows the results from the classification report after evaluation. The Logistic Regression model achieved an accuracy of 83.87% and a macro F1 score of 0.74. The positive sentiment had the highest F1 score at 0.92 and precision of 0.97. The negative class performed well with a precision of 0.70 and high recall of 0.82. However, the model struggled with the neutral sentiment class since this presented the lowest scores with a precision of 0.47 and F1 score of 0.56.

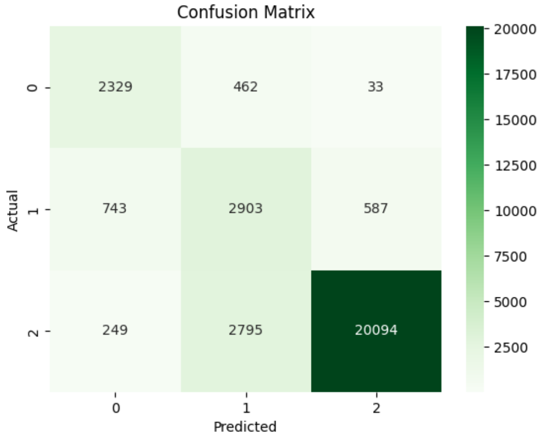

Fig. 3. Confusion matrix.

In Fig. 3, the confusion matrix revealed that the neutral sentiment class was frequently misclassified as negative or positive.

B. BiLSTM

TABLE III
VALIDATION RESULTS

Epoch	Average Loss	Macro F1 Score
1	0.36	0.77
2	0.26	0.78
3	0.22	0.77
4	0.17	0.75
5	0.13	0.75

The BiLSTM model was trained for 5 epochs for 15 minutes. Per Table III, the average training loss decreased across epochs, with the highest macro F1 score being 0.78 on epoch 2, meaning this was saved as the best model.

Table IV shows the classification report when evaluating the model. The BiLSTM model achieved an accuracy of 88.77% and a macro F1 score of 0.78. The positive sentiment had the highest F1 score and recall at 0.95. The negative sentiment class had a high recall of 0.83 and F1 score of 0.79. The neutral sentiment class had a lower recall of 0.57 and F1 score of 0.60.

TABLE IV
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score
Negative (0)	0.75	0.83	0.79
Neutral (1)	0.63	0.57	0.60
Positive (2)	0.95	0.95	0.95
Accuracy			0.89
Macro Avg	0.78	0.78	0.78
Weighted Avg	0.89	0.89	0.89

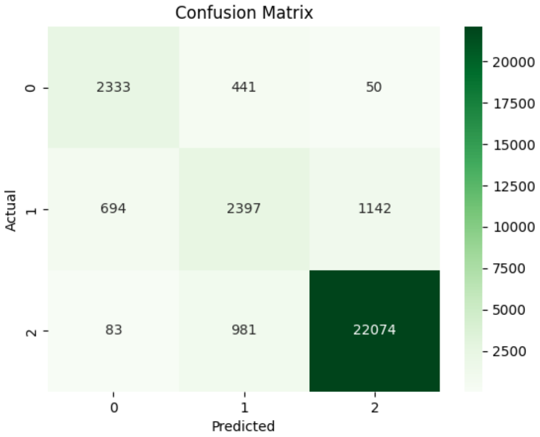

Fig. 4. Confusion matrix.

In Fig 4, similar to the results of the previous model, the confusion matrix revealed the neutral sentiment was often misclassified as positive.

C. DistilBERT

After first running the DistilBERT model with a learning rate of 1×10^{-3} for 5 epochs, the F1 score per class showed that the model was heavily biased towards positive sentiment. Therefore, I decreased the learning rate to 2×10^{-5}, which resulted in an improvement in the F1 scores. Since the F1 score did not show a high improvement from epochs 2 to 5 and this model was computationally heavy, the number of epochs was reduced to 3.

TABLE V
VALIDATION RESULTS

Epoch	Macro F1-Score	Accuracy
1	0.797	0.901
2	0.809	0.906
3	0.807	0.905

After training and evaluating the model at each epoch, requiring approximately 1.22 hours, Table V shows the best model per macro F1 score was the model on the second epoch.

TABLE VI
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score
Negative (0)	0.83	0.78	0.81
Neutral (1)	0.68	0.63	0.65
Positive (2)	0.95	0.97	0.96
Accuracy			0.90
Macro Avg	0.82	0.79	0.81
Weighted Avg	0.90	0.90	0.90

Table VI shows the classification report after evaluation, achieving an accuracy of 90.41%. It shows that the positive sentiment had the best performance with a recall of 0.97 and a F1 score of 0.96. This was followed by the negative class having a high precision of 0.83 and F1 score of 0.81. The neutral sentiment had the lowest recall of 0.63 and F1 score of 0.65, but it still showed an improvement compared to the previous two models.

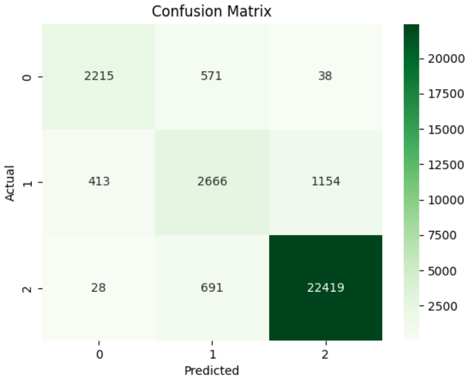

Fig. 5. Confusion matrix.

Fig 5 also reveals that the model still struggled to correctly classify the neutral class, but it showed improvement compared to the results of the previous two models.

D. Summary of Results

All the models performed best on the positive class, which constituted the majority of the dataset. Even though it was a minority class, overall the models performed well on the negative class. However, the neutral class presented the greatest challenge since it was often misclassified.

TABLE VII
MODEL PERFORMANCE COMPARISON

Metric	Logistic Regression	BiLSTM	DistilBERT
Accuracy	83.87%	88.77%	90.41%
Macro F1 Score	0.74	0.78	0.81

The DistilBERT model had the best overall performance, achieving a 90.41% accuracy and a macro F1 score of 0.81. This model had the highest F1 scores per class, notably showing an improvement in the neutral class with an F1 score of 0.65.

While DistilBERT had the best performance, BiLSTM had close accuracy and macro F1 scores to DistilBERT, and Logistic Regression had reasonable accuracy and macro F1 scores. However, Logistic Regression did struggle the most with the neutral class with a low precision of 0.47 and F1 score of 0.56.

In terms of training time, the DistilBERT model had a much higher training time compared to the other models.

Based on these results, using a pre-trained transformer model can improve the accuracy and macro F1 score of sentiment classification when dealing with imbalanced classes.

IV. CONCLUSION

Sentiment analysis can be performed on hotel reviews to assess the guest experience. To classify these reviews as negative, neutral or positive sentiment, I implemented a Logistic Regression model, a BiLSTM model and a pre-trained DistilBERT model.

The results revealed that the pre-trained DistilBERT model had the best performance since it had the highest accuracy and macro F1 score, meaning that using a pre-trained transformer model can enhance sentiment classification tasks. In addition, it showed the best performance when classifying the neutral sentiment. However, despite having the best performance, it was also the most resource-intensive, indicating that depending on the resources available, this model may not be the best. Therefore, if there are limited computing resources available, BiLSTM could be used, which was the next best model with an accuracy and macro F1 score not too different from DistilBERT, while still requiring less training time.

For future work on this dataset, since the dataset contained columns with ratings of other aspects of the hotel, Aspect-Based Sentiment Analysis could be performed.

REFERENCES

[1] M. Sanwal and M. M. Mazhar, “Performance Comparison of Machine Learning and Deep Learning Models for Sentiment Analysis of Hotel Reviews,” *Int. J. Inf. Technol. Appl. Sci.*, vol. 5, no. 1, pp. 1–8, Aug. 2023. [Online]. Available: https://www.researchgate.net/publication/372983916_Performance_Comparison_of_Machine_Learning_and_Deep_Learning_Models_for_Sentiment_Analysis_of_Hotel_Reviews

[2] A. Ameur, S. Hamdi, and S. Ben Yahia, “Sentiment Analysis for Hotel Reviews: A Systematic Literature Review,” ACM Computing Surveys, vol. 55, no. 10, pp. 1—36, Oct. 2023. [Online]. Available: https://www.researchgate.net/publication/372794899_Sentiment_Analysis_for_Hotel_Reviews_A_Systematic_Literature_Review

[3] A. Taparia, “Bidirectional LSTM in NLP,” GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/nlp/bidirectional-lstm-in-nlp/ (accessed Nov. 18, 2025).

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ”DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019. [Online]. Available: https: //arxiv.org/abs/1910.01108