SENTIMENT ANALYSIS OF TRIPADVISOR REVIEWS

Arja HOJNIK, Nives HÜLL, Meta KOKALJ, Ela NOVAK, Katarina PLEMENTAŠ

The project Sentiment Analysis of TripAdvisor Reviews focuses on sentiment analysis of TripAdvisor reviews. It consists of two parts, a dataset creation of 125 TripAdvisor reviews of Michelin-starred restaurants, and the training of seven different NLP models (Logistic Regression, GRU, LSTM, BiLSTM, DistilBERT, BERT Multilingual, TinyBERT, RoBERTa) on a larger dataset. This phase is aimed at enhancing the accuracy and effectiveness of sentiment analysis. The project includes the datasets, annotation guidelines, and trained models, and provides a demonstration of their performance through a demo. The analysis of the models' results provides insights into customer opinions, demonstrating the project's effectiveness in understanding customer feedback.

Keywords: sentiment analysis, annotation, logistic regression, deep learning models, transformer models

1 INTRODUCTION

The project Sentiment Analysis of TripAdvisor Reviews undertaken by Digital Linguistics students at the University of Zagreb as part of a Natural Language Processing (NLP) course, focuses on sentiment analysis of TripAdvisor reviews of Michelin-starred restaurants. Its objectives are to develop a reliable and unbiased dataset for sentiment analysis and to train different NLP models on a larger dataset to enhance sentiment analysis accuracy.

The project consists of two parts, the first part focusing on the creation of a dataset of 125 TripAdvisor reviews of Michelin-starred restaurants, while the second part focuses on training NLP models on a larger dataset of hotel reviews and ratings collected from TripAdvisor, as published on the platform Hugging Face by Niimi Junichiro (Junichiro 2024).

The data acquisition steps included the criteria setting, with a focus on Michelin-starred restaurants noted for sustainability, data extraction, using a

Python script[1] to extract data from the official Michelin Guide website (Michelin Guide 2023), sample selection, using a Python script to randomly select 50 restaurants, and review collection, via a TripAdvisor scraper extension (TripAdvisor® Review Scraper 2023). The result was a dataset comprising of 125 different reviews.

In the second part of the project, seven different NLP models (Logistic Regression, GRU, LSTM, BiLSTM, DistilBERT, BERT Multilingual, TinyBERT, and RoBERTa) were trained and tested on a larger dataset created by Niimi Junichiro (Junichiro 2024). The trained models were then uploaded to Hugging Face and merged into a demo.

The project aimed to evaluate sentiment analysis models by analyzing their predictions. Specifically, the goals were:

1. Assess how well the models perform within the domain they were trained on, focusing on their ability to capture sentiments in hotel reviews.
2. Analyze incorrect predictions across datasets to identify patterns, uncover domain-specific nuances, and inform future improvements in model development and dataset design.
3. Learn and explore various modeling techniques.

The project addressed the following research questions:

1. How well do sentiment analysis models perform when tested on the same domain they were trained on (hotels)?
2. What patterns can be identified in the incorrect predictions across hotel and restaurant datasets?
3. How can the insights gained from prediction analysis guide future model fine-tuning, training strategies, and domain-specific adaptations?

[1] All scripts are available on a GitHub repository: https://github.com/FFZG-NLP-2024/TripAdvisor-Sentiment.

2 LITERATURE REVIEW

When researching the topic of sentiment analysis, we analysed various scientific articles and publications. In the following section, the core theoretical findings from them are presented in a concise manner.

Sentiment analysis, also known as opinion mining, is a domain within natural language processing (NLP). It focuses on identifying and categorizing different opinions, as they are expressed in a specific text to determine attitudes toward specific topics or overall sentiment orientation of a text, with automatic extraction of subjective information from a text and its classification into categories such as positive, negative, and neutral (Tan, Lee, and Lim 2023; Sharma, Ali, and Kabir 2024).

There are many different approaches to sentiment analysis and in this project, we primarily focused on deep learning models and transformer models. One of the most relevant deep learning methods in the field of sentiment analysis are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). They can process text data in a more nuanced way, as they are able to capture contextual and sequential information (Rangarjan et al. 2024). On the other hand, transformer models like BERT (bidirectional encoder representations from transformers) can be used to achieve a higher degree of accuracy within sentiment analysis. This is due to their deep contextual understanding of text, enabled by considering both left and right context of each word in a text (Sharma, Ali, and Kabir 2024).

Sentiment analysis can be performed at various levels of granularity. This can range from individual words or phrases to entire documents. Depending on the task, the analysis can be tailored to best suit its specific needs, such as understanding of a product feature or the general sentiment of a review (Chifu and Fournier 2023; Rangarjan et al. 2024). In our project, we performed sentiment analysis on both sentence-level and review-level.

Employing a range of metrics can be beneficial for sentiment analysis, as this results in a more comprehensive evaluation of model performance. The most common metrics include precision, recall, accuracy, and F1-score. To train robust models that are effective across various fields, it is also important to use diverse and representative datasets (Bharadwaj 2023; Tan, Lee, and Lim

2023).

3 DATASETS

Since including reviews of random restaurants in Europe would be too broad, we decided to narrow our focus on Michelin-starred restaurants only. The constraint we were faced with was the fact the Michelin Guide's website (from which we scraped the restaurants) only allows filtering by cuisine and not by location. Therefore, the project was adjusted to include all restaurants listed on the site, focusing only on those with the "green star" tag.

The second dataset that we worked with was the TripAdvisor Review Ratings dataset, used as a training dataset for our models. It emerged as the best fit for our project since it contains reviews rated on a 5-point scale, which directly matches our test dataset's structure. The data is well-structured and balanced, and its source alignment with TripAdvisor ensures minimal domain drift, as both the training and test datasets share linguistic patterns specific to the broader category of hospitality and user-generated reviews. One limitation of the TripAdvisor dataset is that it primarily consists of reviews for hotels, whereas our test dataset contains reviews for restaurants. Furthermore, its high-quality documentation and structure made preprocessing and integration straightforward. By selecting a dataset that closely matched our test data's format and purpose, we ensured minimal performance degradation and better model generalization within the hospitality domain.

The selected dataset consists of 201,295 rows and includes several columns: 'hotel_id', 'user_id', 'title', 'text', 'overall', 'cleanliness', 'value', 'location', 'rooms', 'sleep_quality', 'stay_year', 'post_date', 'freq', 'review', 'char', and 'lang'. These columns capture various aspects of the hotel experience, with ratings provided for specific categories such as 'cleanliness', 'value', 'location', 'rooms', 'sleep_quality', and 'stay_year'. However, for the purpose of sentiment analysis, we chose to focus solely on the 'text' and 'overall' columns. The 'text' column contains the review content without the title of the relevant review, while the 'overall' column provides an aggregated rating that reflects the user's general sentiment about the hotel. The 'overall' column was chosen as the target variable because it provides a general sentiment about the hotel, which is the primary focus of sentiment analysis.

The statistical report for the 'overall' ratings in the TripAdvisor dataset revealed an uneven distribution, with the majority of reviews being labeled with a rating of 5 (42.98%), while a rating of 1 appeared only in 4%. Considering this, we balanced the dataset, ensuring that each label was equally represented in the final dataset, which was further split into a training set (30,400 reviews), a validation set (1,600 reviews), and a test set (8,000 reviews).

4 METHODOLOGY

4.1 Annotation Dataset Creation

After selecting criteria for scraping, we deployed a Python script (1_get_restaurants.py) to automate the process of restaurant data collection. It utilizes the BeautifulSoup Python library to parse HTML content, extracting relevant data, such as restaurant names, countries, and regions they are located in, and the number of Michelin stars. The script also removes any duplicate entries, therefore ensuring the uniqueness of the dataset, and then compiles the cleaned data into a .txt file.

The script (2_restaurant_selection.py) reads the data from the previously generated .txt file and categorizes the entries by country. It then calculates the proportion of restaurants per country to ensure a representative sample across different regions. Following this, it randomly selects 50 restaurants while maintaining this proportional distribution. The script x_assign_annotator.py then shuffles the list of selected restaurants and distributes them among the five project team members. Each member is therefore assigned 10 unique restaurants. It then outputs the assignments into a .txt file.

Each team member manually verified their assigned restaurants on TripAdvisor, confirming their presence on the website and checking the number of reviews written in English. In cases where restaurants had fewer reviews than needed (25), a decision was made to retain them, in order to preserve the dataset's representativeness. We then used the Chromium-based web extension TripAdvisor® Review Scraper to scrape reviews from each restaurant on TripAdvisor. A script 3_clear_reviews.py was deployed after scraping the reviews to organize and clean them, ensuring that the data is formatted properly.

The script 4_sample_reviews.py was deployed to process restaurant reviews and prepare a test dataset of 25 random stratified reviews. It ensured a uniform length of reviews, indexed them, and saved appropriate reviews into a .csv file. The script was later deployed again with changed parameters to include 125 reviews instead of 25. Following the collection, the reviews underwent a preliminary sentiment analysis using the 7_preannotate_sentiment.py script, which assigned sentiments to the reviews to allow a swifter manual annotation.

4.2 Annotation

The first version of our annotation guidelines utilized a rating scale ranging from 1 to 5. The determined scores were the following: 1 (negative), 2 (mostly negative), 3 (neutral), 4 (mostly positive), and 5 (positive).

25 reviews collected with the script 4_sample_reviews.py were annotated by all five team members and interannotator agreement was calculated. Ambiguous cases were resolved and a second version of the annotation guidelines was created based on feedback of the test annotation. It included minimal changes, especially relating to resolving reviews, where multiple sentiments were expressed.

After collecting 125 using the same script, a decision was made for each review to be annotated by three different annotators (and not all five). Each annotator therefore annotated 75 reviews. However, these reviews were then split into individual sentences and the reviews were annotated on both sentence- and review-levels. The final version of annotated samples was analyzed using aggregated agreement and assigned a final label.

To calculate the aggregated agreement and the final label for the review paragraphs in our dataset, we used Excel. The reviews in our dataset were divided into sentences, and each sentence was annotated by multiple annotators. However, we decided to calculate one unified label for the entire review (paragraph) rather than keeping the labels for individual sentences, meaning that for every review, we treated the paragraph as one single unit and calculated an overall score based on the labels provided by the annotators.

The process started with manually inputting the values from the annotators' individual annotation sheets into a new aggregated Excel sheet. This allowed us to have a centralized file where all three annotators' scores for each review were collected and processed for calculating the unified label. For each review, we wrote the index of the review (corresponding to its original position in the dataset) in the first column. Then, we transferred the paragraph-level scores given by the three annotators from their individual Excel sheets into the corresponding columns of the new sheet. We only worked with the scores that were provided by the annotators.

To calculate the aggregated score, we used the formula =AVERAGE() in Excel, which takes the values provided by the annotators and calculates the mean of these values. After calculating the average score, we used the formula =ROUND() to round the result to the nearest whole number, as the final label needed to be an integer. The rounded result would be the final label for that review.

4.3 Modeling Dataset Creation

From all of the columns in the TripAdvisor Review Ratings dataset, we chose to focus solely on the 'text' (containing the review content) and 'overall' (providing aggregated rating) columns, which were extracted from the loaded dataset. The 'overall' column was renamed to 'label' to clearly identify it as the target variable for the sentiment analysis.

To address the imbalance of the chosen dataset, we first created a smaller sampled dataset to ensure that each label is adequately represented. We noticed that the smallest group of reviews had just over 8,000 instances. To ensure an equitable representation across all labels, we selected a cut-off of 8,000 reviews per label for sampling.

We thus balanced the dataset by randomly sampling 8,000 reviews from each label, ensuring that no label dominates the training data. This gave us a total of 40,000 reviews (8,000 reviews x 5 ratings), ensuring that each label was equally represented in the final dataset.

Once the dataset was balanced, we proceeded to split it into training, validation, and test sets. The data was first divided into an 80-20 split, where 80% of the data was reserved for training and validation, and the remaining 20% was used for testing. The split is done using the train_test_split function from the sklearn.model_selection module, with the stratify parameter to ensure that the label distribution remains consistent across both parts. The training and validation data were further split into 75% for training and 25% for validation. Again, stratified sampling is used to maintain the label distribution. The training set contains 30,400 reviews, the validation set has 16,000 reviews, and the test set includes 8,000 reviews.

After splitting, the three datasets (train, validation, and test) were saved as Parquet files. The training set was saved as a train.parquet, the validation set as validation.parquet, and the test set as test.parquet. Furthermore, to ensure broader availability, the dataset was pushed to the Hugging Face Hub, making it accessible to others under the repository name nhull/tripadvisor-split-dataset-v2.

4.5 Machine Learning

The first method we tested for sentiment analysis was Logistic Regression. Using our TripAdvisor review dataset, we vectorized the review text using TF-IDF with n-grams (unigrams, bigrams, and trigrams) and a maximum of 10,000 features. After hyperparameter tuning with GridSearchCV (testing regularization strengths C of 0.1, 1, 10, and 100), the best-performing model was trained and evaluated.

4.6 Deep Learning

In our study, we evaluated three deep learning architectures—LSTM, GRU, and BiLSTM—on the TripAdvisor dataset. We focus on one-layer architecture for each model, leveraging pre-trained GloVe embeddings, as we aim to provide semantically rich word representations to our models.

The TripAdvisor dataset was preprocessed to convert the reviews to lowercase, and non-alphabetic characters were removed using regular expressions. Sentiment labels were normalized to five classes, ranging from 0 (negative) to 4 (positive) to align with zero-based indexing in machine learning models. Text was tokenized into sequences using a maximum vocabulary size of 10,000 words, and sequences were padded to a fixed length of 200 for uniformity across samples. Pre-trained 100-dimensional GloVe embeddings (glove.6B.100d.txt) were used to initialize the embedding layer for each model.

We employed Keras Tuner's Hyperband algorithm to optimize the following hyperparameters:

- Number of units in each recurrent layer (64–256 in steps of 64).
- Dropout rate (0.2–0.5 in steps of 0.1).
- Number of dense units in the fully connected layer (32–128 in steps of 32).
- Optimizer choice (adam or rmsprop).

Each model was partially trained for up to 10 epochs with early stopping to determine the best hyperparameters. The best configurations were used for training the models for up to 20 epochs with a patience of 3 epochs to avoid overfitting. Models were evaluated on the Tripadvisor test set to compare performance.

After completing training and validation, we evaluated the models on the TripAdvisor test dataset, which consists of 8,000 reviews. With GRU identified as the best-performing architecture, we evaluated its performance on the unseen Michelin dataset, to analyze the cross-domain generalization ability of our models. The Michelin dataset, composed of reviews for restaurants, provided a distinct domain compared to the hotel-focused TripAdvisor dataset.

4.7 Transformers

4.7.1 DistilBERT

The base model, DistilBERT (distilbert-base-uncased), initially achieved a testing accuracy of approximately 60% when evaluated on our TripAdvisor review dataset. To improve its performance, we fine-tuned the model using a learning rate of 3e-05, a batch size of 64, and early stopping over 10 epochs (patience set to 5). Additionally, we experimented with hyperparameter optimization strategies, including grid search and Optuna, but these did not result in significant performance improvements.

4.7.2 TinyBERT

The TinyBERT model was also trained and evaluated on the TripAdvisor sentiment analysis dataset, with the objective of fine-tuning the pre-trained language model to classify text into five sentiment categories.

Two versions of the TinyBERT model were trained using slightly different configurations to optimize performance. The methodology for both models involved tokenizing the dataset using the TinyBERT tokenizer, with truncation and padding applied to a maximum sequence length of 128 tokens. The huawei-noah/TinyBERT_General_4L_312D was used as the pre-trained model base and fine-tuned for the 5-class sentiment classification task.

The first version, TinyBERT1, was fine-tuned with the following configuration: a learning rate of 2e-5, a batch size of 16 (used for both training and evaluation), 3 training epochs, weight decay of 0.01, and evaluation conducted at the end of each epoch.

TinyBERT2, differed slightly, utilizing a learning rate of 1e-5, a batch size of 32 (for both training and evaluation), 5 training epochs, weight decay of 0.005, and an evaluation strategy where performance was assessed every 500 steps. The key differences between the two setups were the evaluation frequency, learning rate, and batch size, which influenced the model's generalization ability and stability during training.

4.7.3 RoBERTa

Like other models, RoBERTa was trained on TripAdvisor hotel reviews and was later additionally tested on 125 TripAdvisor restaurant reviews to explore the model's generalization to a related but different domain. The text was tokenized with a maximal length of 128.

The training configuration for the RoBERTa model included a learning rate of 5e-5, a batch size of 16 for the hotel review dataset, and a batch size of 8 for the restaurant review dataset. The training was conducted over 3 epochs using the AdamW optimizer with a weight decay of 0.01. A linear scheduler with a warmup phase was applied to adjust the learning rate during training.

5 RESULTS

5.1 Machine Learning

The evaluation of the model’s performance provided the following results:

- **Testing Accuracy:** 61.05%
- **Macro F1-Score:** 0.6088
- **Weighted F1-Score:** 0.6088
- **Macro Average Precision:** 0.6077

The model performed well for extreme sentiment labels (1 and 5), with precision values of 0.6989 and 0.7053, respectively. However, it struggled with mid-range labels (2, 3, and 4), where precision and recall were notably lower. This performance disparity is visualized in the confusion matrix in the appendix (Image 1).

The Logistic Regression model performs well on extreme sentiment labels (1 and 5) with high true positive counts, while it frequently misclassifies mid-range labels (2, 3, and 4), often predicting adjacent sentiment levels. While this model provided a simple and interpretable baseline, its limitations in capturing nuanced sentiment distinctions, particularly for mid-range labels, highlighted the need for more advanced methods.

We also experimented with a Random Forest classifier using its default settings as an additional baseline. However, it demonstrated lower performance, with accuracy below 60%, and thus was not pursued further for this task.

In comparison to the fine-tuned DistilBERT model, Logistic Regression performed slightly worse in terms of accuracy and F1-scores but served as an essential starting point for this analysis.

5.2 Deep Learning

From the results of the evaluation on the TripAdvisor test dataset, GRU emerged as the best-performing model, achieving the highest accuracy (62.16%), with an F1-score (0.62) comparable to BiLSTM. The LSTM model performed slightly worse, while BiLSTM offered comparable results but with slightly higher computational cost. The GRU model demonstrated better generalization due to its simplicity and efficiency, making it more effective at capturing sequential patterns in the dataset.

Model	Accuracy (%)	Precision	Recall	F1-Score
LSTM	60.41	0.60	0.60	0.60
GRU	**62.16**	0.62	0.62	0.62
BiLSTM	61.67	0.62	0.62	0.62

Table 1: Deep Learning Models' Evaluation Scores.

The confusion matrices for the BiLSTM, GRU, and LSTM models (see Images 2–4 in the appendix) demonstrate varying levels of performance across the five classes. The BiLSTM model achieves relatively strong diagonal dominance, indicating good accuracy in correctly classifying most labels, especially for classes 1 and 5. However, some confusion arises between neighboring classes, such as class 2 being misclassified as class 3 or vice versa. The GRU model shows similar patterns but with slightly more confusion between classes, especially between class 3 and classes 4 and 5, which indicates lower precision and recall for those classes compared to BiLSTM. Lastly, the LSTM model also captures the correct labels reasonably well, but there is more spread in the off-diagonal values, particularly for class 3, which often gets misclassified as class 4, and vice versa.

5.2.1 Testing on the Michelin Dataset

After selecting GRU as the best-performing architecture, we evaluated its performance on the unseen Michelin dataset. The results are as follows:

Dataset	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Michelin	**68.28**	**0.68**	**0.68**	**0.68**

Table 2: GRU Evaluation Scores on the Michelin Dataset.

The GRU model achieved 68.28% accuracy and an F1-score of 0.68 on the Michelin dataset, demonstrating robust performance when applied to a domain distinct from the training data. Notably, the test set results slightly outperformed the validation set predictions, a phenomenon that is uncommon in machine learning. This could be attributed to the Michelin dataset's review style or language being more predictable and better aligned with the patterns learned during training.

The GRU model performs reasonably well, particularly in distinguishing classes 1 and 5, but its performance decreases when dealing with overlapping or neighboring class distributions. This suggests that while the GRU effectively captures certain patterns, further refinement or feature engineering may be needed to reduce inter-class confusion.

5.3 Transformers

5.3.1 DistilBERT

The fine-tuned model, nhull/distilbert-sentiment-model, achieved the following performance:

- **Testing Accuracy:** 63.91%

- **Macro F1-Score:** 0.64
- **Weighted F1-Score:** 0.64
- **Macro Average Precision:** 0.6140
- **Bias:** The model overpredicts sentiment by an average of 0.39.

While precision and recall were higher for extreme sentiment labels (1 and 5), performance for mid-range labels (2 and 3) remained weaker, reflecting some limitations in capturing subtler sentiment distinctions. The confusion matrix for the fine-tuned DistilBERT model is shown below, highlighting the model's strengths in predicting extreme sentiment classes and its struggles with mid-range labels (see Image 6 in the appendix).

To investigate whether the issue was due to the 5-level sentiment classification setup or the domain specificity of the dataset, we also tested nlptown/bert-base-multilingual-uncased-sentiment, a pre-trained multilingual model designed for 5-star sentiment classification. Without additional fine-tuning, this model achieved a slightly lower testing accuracy of 61.46% on the same dataset and exhibited a similar overprediction bias with an average error of 0.44. Like the fine-tuned DistilBERT model, its performance on mid-range sentiment labels was inconsistent. The confusion matrix for this multilingual model is shown below, illustrating similar patterns of strong performance on extreme sentiment labels and weaker results for mid-range sentiment (see Image 7 in the appendix).

Fine-tuning DistilBERT provided a modest improvement over the base model and outperformed the multilingual model. However, the dataset's subjectivity and reliance on overall ratings, which do not always reflect the full content of reviews, likely influenced the results.

5.3.2 TinyBERT

The evaluation of both TinyBERT models showed the second variation of the model to be performing the best, with an accuracy of 0.6535. It was then also evaluated based on its precision, recall, and F1 score. The confusion matrix for the second training is shown in Image 8 in the appendix.

The results of the TinyBERT2 model were as follows:

- **Accuracy**: 0.6535
- **Precision**: 0.635
- **Recall**: 0.641
- **F1-Score**: 0.636

The adjustments in the training parameters between the versions impacted the models' ability to fine-tune effectively. The second strategy of evaluating every 500 steps provided more frequent feedback and potentially more stable learning outcomes, as seen in its superior performance metrics compared to the first version.

The results indicate that while TinyBERT models are capable of handling complex sentiment analysis tasks, there is room for improvement, particularly in distinguishing between closely related sentiment classes as seen in the confusion matrix.

5.3.3 RoBERTa

The model's loss decreased from 0.8893 in the first epoch to 0.5313 in the third epoch, indicating effective learning. The model achieved an accuracy of 65.94% on the hotel validation dataset. Precision, recall, and F1-score were generally consistent across classes, with class 5 performing best.

On the hotel test dataset, the model achieved an accuracy of 67.22%, with class 5 again performing the best. Despite being trained on hotel data, the model achieved an accuracy of 74.40% on the restaurant dataset. However, performance varied significantly across classes, with class 5 performing best.

The results of the RoBERTA model for hotels were as follows[2]:

- **Accuracy**: 0.672
- **Precision**: 0.674
- **Recall**: 0.672

[2] See cofusion matrix in Image 9 in the appendix.

- **F1 Score**: 0.674

The results of the RoBERTA model for restaurants were as follows[3]:

- **Accuracy**: 0.744
- **Precision**: 0.81
- **Recall**: 0.74
- **F1 Score**: 0.73

7 DISCUSSION

7.1 Machine Learning

Even though the Logistic Regression model performed well for extreme sentiment labels (1 and 5), it struggled with mid-range labels (2, 3, and 4). This limitation is likely due to the model's reliance on individual words or phrases, which makes it less effective at capturing nuanced sentiments. The confusion matrix revealed frequent misclassifications between adjacent sentiment levels, particularly for mid-range labels, which suggests that Logistic Regression is not well-suited for capturing the subtleties of sentiment expression, especially in reviews that contain mixed or complex sentiments.

7.2 Deep Learning

While the GRU model outperformed its counterparts, achieving moderate success, these results remain below the benchmarks seen in sentiment analysis tasks with high-quality datasets (often reaching 80% or higher). The relatively modest scores (accuracy in the 60–68% range) suggest challenges inherent to the dataset, such as:

- Ambiguity in Mid-Range Labels (2 and 3): These labels involve subtle language cues, making them more difficult to classify accurately.

[3] See cofusion matrix in Image 10 in the appendix.

- Domain-Specific Challenges: Differences between the TripAdvisor and Michelin datasets may have impacted generalization.
- Limited Training Dataset Size: While sufficient, the dataset could benefit from augmentation or expansion.

The slight improvement on the Michelin dataset compared to the TripAdvisor test set is noteworthy, as cross-domain generalization often results in lower performance. This could suggest that the Michelin dataset contains more predictable patterns or aligns well with the training data.

Although the results obtained were reasonable given the dataset's characteristics, several strategies could be explored to improve performance:

1. Augment the Dataset: Techniques such as paraphrasing or synonym replacement could expand the training dataset and improve the model's ability to handle diverse linguistic styles.
2. Exhaustive Hyperparameter Tuning: A broader search over hyperparameter space, including longer training epochs, could uncover configurations yielding higher accuracy and F1-scores.
3. Class Balancing: Oversampling underrepresented classes or undersampling overrepresented ones could mitigate performance discrepancies across sentiment labels.
4. Ensemble Methods: Combining predictions from multiple models or experimenting with deeper architectures could lead to performance improvements by leveraging complementary strengths.

7.3 Transformers

The **DistilBERT** model, after fine-tuning, achieved a testing accuracy of 63.91%, outperforming the baseline Logistic Regression model. However, like the deep learning models, DistilBERT struggled with mid-range sentiment labels, reflecting the challenges of capturing subtle sentiment distinctions. The model also tended to overpredict sentiment, with an average bias of 0.39, suggesting that while transformer models like DistilBERT offer improved accuracy, they may still struggle with the subjectivity and complexity of realworld reviews.

The **TinyBERT** model, in its second iteration, achieved an accuracy of 65.35%, demonstrating the potential of smaller transformer models for sentiment analysis tasks. However, the model's performance was still limited by its ability to distinguish between closely related sentiment classes, as evidenced by the confusion matrix. The adjustments in training parameters, such as evaluation frequency and learning rate, had a noticeable impact on the model's performance, highlighting the importance of hyperparameter tuning in optimizing transformer models.

The **RoBERTa** model achieved an accuracy of 67.22% on the hotel test dataset and 74.40% on the restaurant dataset. While the model performed well on extreme sentiment labels, its performance varied significantly across classes, particularly for mid-range labels.

8 DIFFICULT AND INTERESTING CASES

In the analysis of prediction errors, the Logistic Regression model showed 46 cases where predictions differed by 3 or more, with 33 differing by 3 and 13 by 4. Errors often stemmed from the model's reliance on individual words or phrases, such as in "Started off rough, ending wonderful," where the negative start overshadowed the positive conclusion. Similarly, sarcastic or unconventional language, like "This is a great place to stay if you are visiting the NY parole office or nudie bar next door," misled the model.

DistilBERT exhibited 25 cases with differences greater than 3 in sentiment prediction, including 5 instances with a difference of 4. It struggled with dataset mislabeling, figurative language, mixed sentiments, and verbose or fake reviews, leading to errors in sentiment prediction. TinyBERT[4] faced similar issues, particularly with mixed sentiments, where negative phrases in positive contexts caused misclassification. RoBERTa[5] performed better (with

[4] See Examples 1–3 in the appendix.

[5] See Example 4 in the appendix.

94.62% of incorrect predictions being just one point off), with most errors being one point off, but it still struggled with figurative language, sarcasm, and nuanced sentiments due to over-reliance on sentiment markers and keywords. To improve, models should be fine-tuned with enriched datasets, advanced attention mechanisms, and data augmentation to handle complex linguistic nuances.

The GRU model had 44 cases with errors of 3 or more, including 29 with a difference of 3 and 15 with a difference of 4. It struggled with sarcasm and sudden sentiment shifts, suggesting a need for training on datasets with such examples and a sub-module for sarcasm detection.

The LSTM model also showed 44 errors of 3 or more, with 36 at a difference of 3 and 8 at a difference of 4. Challenges included metaphorical language and mixed sentiments, which could be addressed with multi-task learning and a larger, more diverse corpus.

The BiLSTM model recorded 69 errors of 3 or more, with 53 at a difference of 3 and 16 at a difference of 4. Errors stemmed from reliance on individual words or phrases, which could be improved by dependency parsing for better context and sentiment-specific word embeddings for nuanced expression sensitivity.

9 CONCLUSION

The project successfully navigated the complexities of sentiment analysis within the hospitality domain, specifically focusing on Michelin-starred restaurants and hotel reviews. The project achieved its primary objectives of developing a reliable and unbiased dataset and training various NLP models to enhance sentiment analysis accuracy. Sentiment analysis models demonstrated moderate performance within the same domain of hotel reviews. Transformer models, particularly RoBERTa, showed the most promising results, indicating their superior ability to capture contextual nuances compared to traditional machine learning and other deep learning models. Incorrect predictions predominantly arose from the models' inability to effectively interpret nuanced and complex sentiment expressions. The overlap and subtle distinctions between adjacent sentiment classes further complicated accurate classification. Across all models, accurately classifying mid-range sentiments (2, 3, and 4) remained a significant challenge. Common issues included handling sarcasm, mixed sentiments within reviews, and the reliance on individual words or phrases that do not capture the overall sentiment context.

Given the constraints of this project, our experimentation with the restaurant dataset was limited. The small size of the restaurant dataset presented a significant challenge, as it restricted our ability to thoroughly evaluate and optimize the model's performance in this domain. Additionally, the lack of time and resources meant that data augmentation techniques, which could have potentially improved the model's robustness and accuracy, were beyond the scope of this project.

To address these challenges and further enhance the performance of the Roberta model across different domains, several strategies can be considered for future work. Augmenting the training data is crucial for improving robustness. This would help prevent overfitting to the more abundant hotel dataset and ensure that the model learns from a diverse set of examples.

Fine-tuning the model on specific domains, such as restaurants, could significantly enhance its performance by allowing it to capture domain-specific nuances and improve its ability to generalize within those domains. This cross-domain evaluation approach involves adapting the model with domain-specific fine-tuning to optimize performance on specialized datasets.

Furthermore, enhanced evaluation strategies can be implemented to stabilize model training. Introducing evaluation checkpoints during training, such as evaluating the model's performance at regular intervals (e.g., every n-steps), can help monitor its stability and effectiveness throughout the training process. This approach allows for early detection of overfitting or underfitting and facilitates timely adjustments to training parameters.

Additionally, experimenting with different hyperparameters is essential for optimizing the model's overall accuracy and consistency across classes. By systematically exploring various hyperparameter combinations, it is possible to identify settings that improve the model's ability to distinguish between different sentiment levels effectively.

10 REFERENCES

Aliyu, Yusuf et al. 2024. ‘Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources’. *IEEE Access, 12,* 66883–66909. doi: 10.1109/ACCESS.2024.3398635.

Bharadwaj, Lakshay. 2023. ‘Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification’. *International Journal For Multidisciplinary Research* 5 (September). https://doi.org/10.36948/ijfmr.2023.v05i05.6090.

Chifu, Adrian-Gabriel, and Sébastien Fournier. 2023. ‘Sentiment Difficulty in Aspect-Based Sentiment Analysis’. *Mathematics* 11 (November):4647. https://doi.org/10.3390/math11224647.

Ganie, Aadil Gani. 2023. ‘Presence of informal language, such as emoticons, hashtags, and slang, impact the performance of sentiment analysis models on social media text?’. arXiv preprint arXiv:2301.12303. https://arxiv.org/abs/2301.12303

Junichiro, Niimi. 2024. ‘Hotel Review Dataset (English)’. https://github.com/jniimi/tripadvisor_dataset.

Michelin Guide. 2023. ‘Michelin Guide: Official Website’. https://guide.michelin.com.

Rangarjan, Prasanna, Bharathi Mohan Gurusamy, Gayathri Muthurasu, Rithani Mohan, Gundala Pallavi, Sulochana Vijayakumar, and Ali Altalbe. 2024. ‘The Social Media Sentiment Analysis Framework: Deep Learning for Sentiment Analysis on Social Media’. *International Journal of Electrical and Computer Engineering (IJECE)* 14 (June):3394. https://doi.org/10.11591/ijece.v14i3.pp3394-3405.

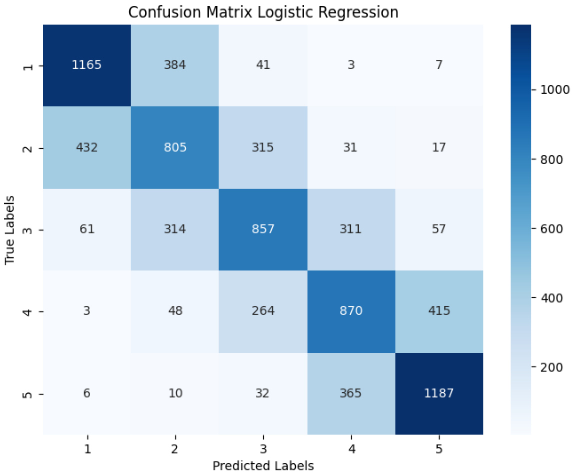
Sharma, Neeraj, A B M Shawkat Ali, and Ashad Kabir. 2024. ‘A Review of Sentiment Analysis: Tasks, Applications, and Deep Learning Techniques’. *International Journal of Data Science and Analytics*, July, 1–38. https://doi.org/10.1007/s41060-024-00594-x.

Tan, Kian Long, Chin Poo Lee, and Kian Ming Lim. 2023. ‘A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research’. *Applied Sciences* 13 (7). https://doi.org/10.3390/app13074550.

TripAdvisor® Review Scraper. 2023. ‘TripAdvisor® Review Scraper Chrome Extension’. https://chromewebstore.google.com/detail/TripAdvisor%C2%AE%20Review%20Scraper/pkbfojcocjkdhlcicpanllbeokhajlme.

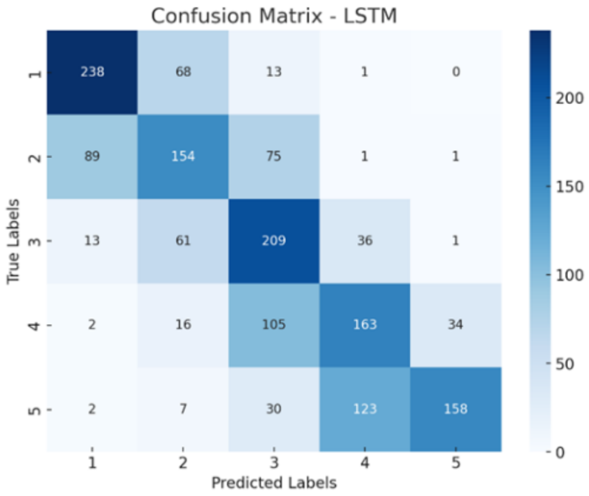
11 APPENDIXES

Appendix A



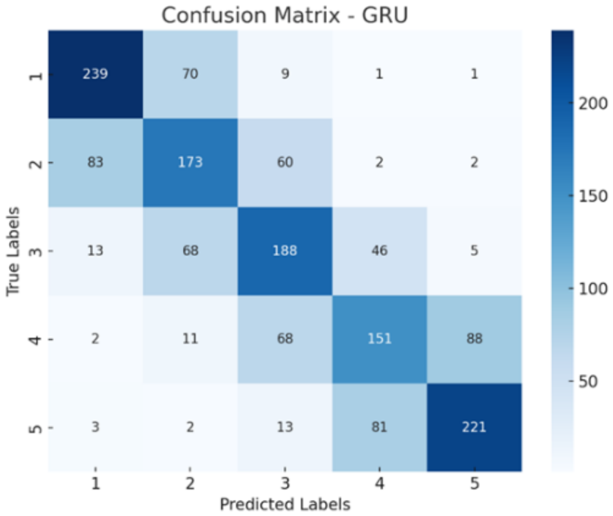
	1	2	3	4	5
1	1165	384	41	3	7
2	432	805	315	31	17
3	61	314	857	311	57
4	3	48	264	870	415
5	6	10	32	365	1187

Image 1: Confusion Matrix for Logistic Regression



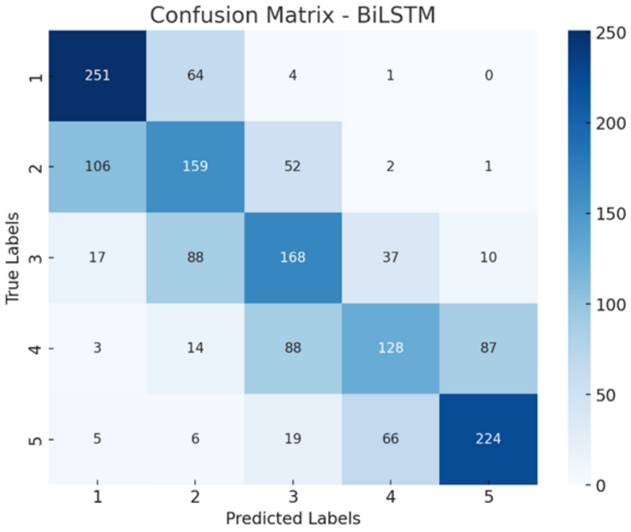
	1	2	3	4	5
1	238	68	13	1	0
2	89	154	75	1	1
3	13	61	209	36	1
4	2	16	105	163	34
5	2	7	30	123	158

Image 2: Confusion Matrix for LSTM



True \ Predicted	1	2	3	4	5
1	239	70	9	1	1
2	83	173	60	2	2
3	13	68	188	46	5
4	2	11	68	151	88
5	3	2	13	81	221

Image 3: Confusion Matrix for GRU



True \ Predicted	1	2	3	4	5
1	251	64	4	1	0
2	106	159	52	2	1
3	17	88	168	37	10
4	3	14	88	128	87
5	5	6	19	66	224

Image 4: Confusion Matrix for BiLSTM

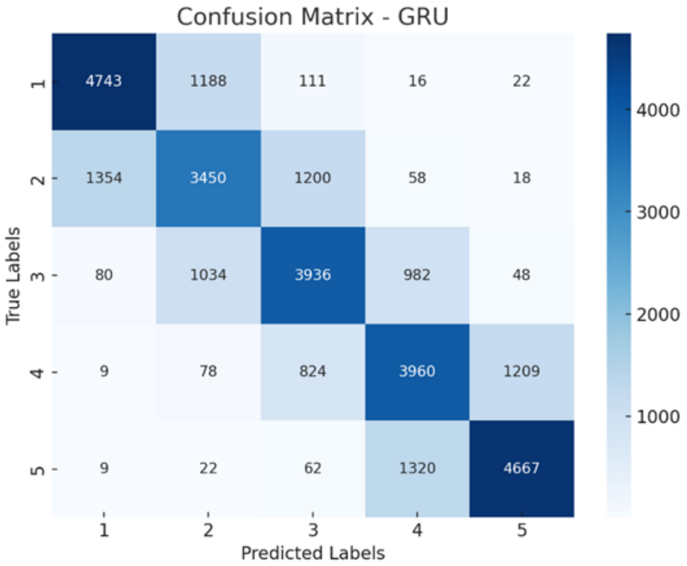


Image 5: Confusion Matrix for GRU (Michelin Dataset)

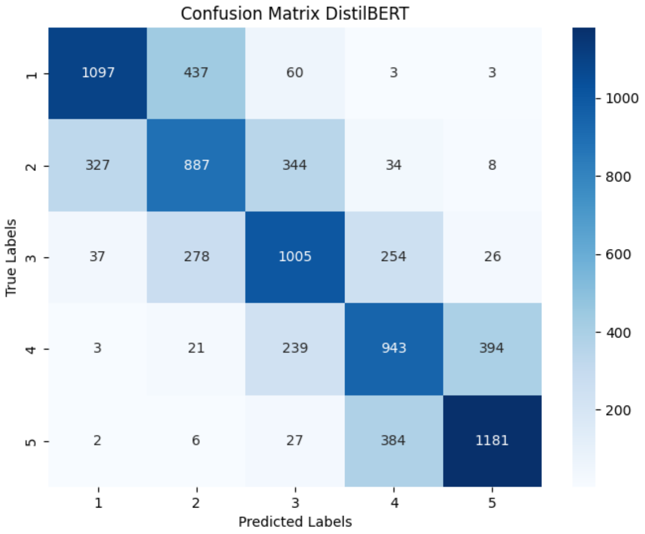


Image 6: Confusion Matrix for DistilBERT

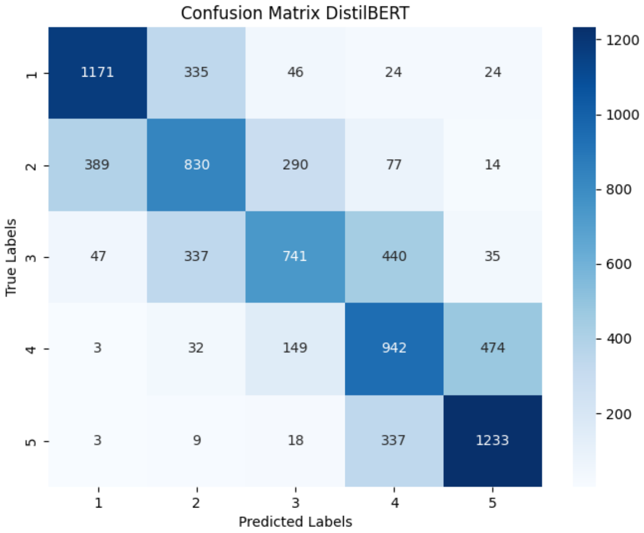


Image 7: Confusion Matrix for Fine-tuned DistilBERT

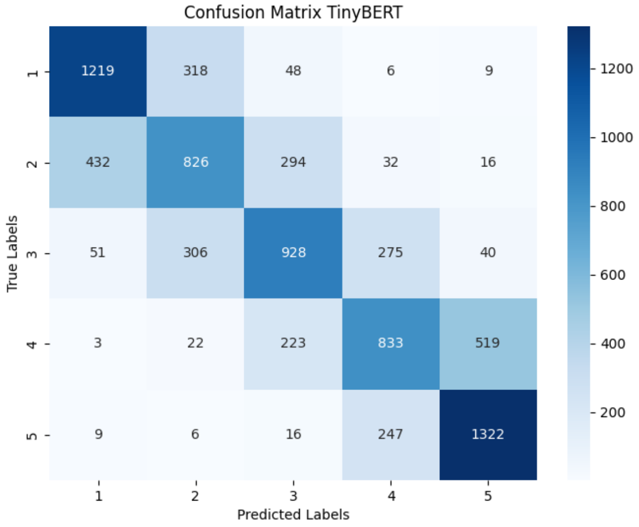


Image 8: Confusion Matrix for TinyBERT

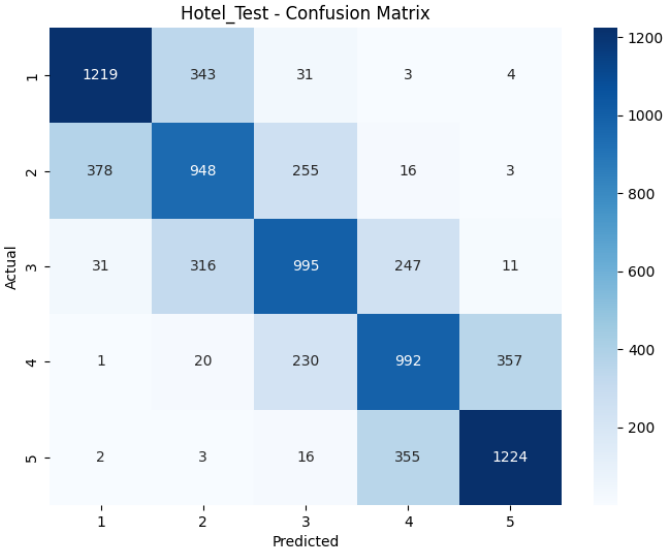


Image 9: Confusion Matrix for RoBERTa

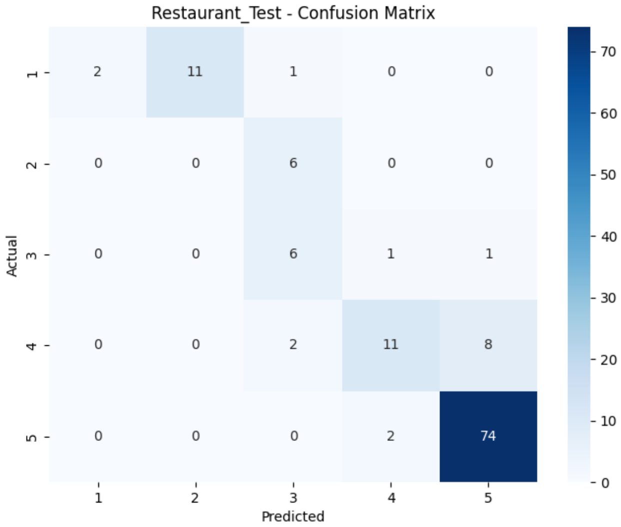


Image 10: Confusion Matrix for RoBERTa (Michelin dataset)

Appendix B

Example 1:
We had a great time at the Days Inn, we did not get a complementary breakfast, but ate at the resturant on site, which could do with a broader offering, not everyone likes bacon.

Predicted Sentiment: 1
Actual Sentiment: 4

Example 2:
This hotel is in a great location for visiting the historical sites in Philadelphia. Th rooms were recently renovated and were beautiful. There are tons of shops, bars and restaurants within walking distance. The area around the hotel was much more quiet than in the Center City area where we spent a night at the Doubletree. I would definitely pick this area over the other areas in the city. Cabs were plentiful and you could go anywhere else in the city for under $10.

Predicted Sentiment: 1
Actual Sentiment: 5

Example 3:
I didn't actually stay here - I live in NYC, but I ran in seeking shelter from the storm one afternoon when I got caught in the rain, and was blown away by the stunning lobby. The staff greeted me warmly. I asked if I could buy an umbrella and they were very accommodating. The umbrella was a little pricey, however after discarding the plastic wrapping I felt the softest, most comfortable umbrella handle I have ever gripped my fingers around. I left the hotel and a gust of wind enveloped me, but didn't blow the umbrella inside out as so often happens. My Paciottis were saved. Thank you Trump Soho! I plan to go back in the summer to check out Bar d'Eau.

Predicted Sentiment: 1
Actual Sentiment: 5

Example 4:

Setai Have Broken my Heart, I didn't Want To Leave...

Continuing celebrating our 30th wedding anniversary, we booked Suite for 3 nights for the price of 2.

Pros:

-Very polite, friendly and helpful staff, will go out of their way to ensure your stay is happy.

-Always address by Mrs... or Mr...

-Will accommodate you with the best that's available if capacity too full

-I like check in and check out any time, but out of consideration for other guests and room service staff, we checked in at 3:00 PM and checked out at 3:00pm again out of consideration for other guests, even though we could have left later and I wanted to. There for since we were busy with other matters, we did everything we planned on our checkout day and were very satisfied to not have missed on anything we planed.

-Non alcoholic drinks in fridge plenty. And what a marvelous idea was to suggest us if we only drink water then they will replenish it only with water, if you need more they will deliver, but we always had enough.

-N espresso is great in the room, I have one home and addicted to it, so if you finish all of the coffee offered, call in the evening and they will deliver you more to have in the morning.

-Quiet rooms so if you meditate, its great/or want to take siesta or fall asleep early, notify front desk not to be disturbed if your sign is on. The only time someone will knock are the fridge/coffee replenish or bring you cookies and will do turn down service if you desire.

-For a suite the accommodation was huge and when the person knocked to deliver ice I was in the bathroom and it was a walk to get to the door, after I opened the door he thought no one was in and was about to leave, I thought that was funny explaining to him that it was a long walk to the door.

-The rooms are well appointed and the decor is pleasant to the eyes. Room and Hotel itself is very clean.

-If you are like me staying in 5 stars hotels the bathroom are pretty much the same in renovated hotels, ours was huge and I loved the window in the bathroom, we've been partying often in there while waiting for my service dog to finish his dinner, he is spoiled and wants us to be around when he its.

-Two TVs are standard in NYC for suits. And if you never sow TV in the mirror, I somehow felt the screen was bigger than in other hotels.

-Bed is comfortable, so are pillows, I absolutely adored the style of the windows, honestly that was my first choice to stay in Setai, provides with ability to look down without stretching.

your neck. And lower window opens from the top which is less scarier than from the bottom as in Trump Soho.

-If after your room was cleaned and you noticed that something wasn't replenished - a phone call away and a minute of wait.

-If your top cover not on the bed with little neck pillow, it is most likely in the closet.

-Loved the Foyer, Powder room with the door.

-Real walk in room closet.

-Bath tub is for two and the biggest I was in aside from that in Mercer.

-Rain Shower.

-the hotel has Souna and Steam room, Gym, with private room for floor exercise alone or with personal trainer.

-If you take advantage of their spa, you also receive access to whirl pool, hamam, and rain shower from all over top, walls (this kind of shower was my first time experience to see) and plenty of healthy snack and fruits and drinks. The spa person will be more than happy to give you the tour. We did the tour of the hotel while waiting for our room cleaned.

-We didn't use the bar or restaurant, because my husband already planned the places to eat ahead of arrival and it is his department. If you are like us who enjoy rather unquiet ambiance the kind that no one can hear your conversation we liked Dallas and Irish Pub a block away. We drink protein shakes for breakfast and lunch, but the day of our check out we walked over to Barking Dog which is famous for their Pancakes which I like and my husband ordered eggs and mixed breakfast meats.plate. Don't order toast, it comes with the good size bun.

-Places to see Bryant Park, always something going on there, there is a watering whole and instead of benches there are chairs and tables. Which are scattered all over and you can find private spot to enjoy and then move to another spot that is what we were doing. I also think the chairs for the reason that a person

don't make a bed out of it. Park is opened until 9PM. Security is present. Oh and they also have ping pong table there to rent.
-Library a must see, I've been there many times when I was studying in University. But being there as a tourist is not the same experience, things I missed. They always have something in the exhibition room that is unique theme.
-AC is quiet and works perfect with 3 room temperature thermostats.
-Not like I ever seen in other hotels the channels managary is organized in sections, where each section of channels is listed in TV guide with consecutive numbers. I don't recall seeing such in other hotels.
Cons: non I can think of. If you are experienced traveler you know what to do, if you are not, don't upset yourself with something that wasn't done the way you expected, just ask they are there to ensure you leave Setai Happy.
-Suites are expensive, but once you try one you cannot go below, but in my opinion there is nothing like staying in the suit. But if you are in New York Sight Seeing and believe me you will be disappointing not booking longer stay, but even if it is longer stay you still will not have enough days to see all. We've been coming to NYC on vacation 12 times and still didn't see everything we wanted to see. But since we are local we mostly come to enjoy the room and the view is the top priority for me.
-I wish we could have stayed on the higher floor that would really complete our joy.
From my experience if you come with the right attitude you will do your self and your health a lot of good and will only bring good memories back. This worked in every hotel in NYC we stayed at, except Soho Trump with the staff there Guaranteed to ruin not just your stay but entire vacation, and you will stay 5 categories below of the room you reserved in SohoTrum for the price double of what you pay in Setai.

Predicted Sentiment: 1
Actual Sentiment: 5

Possible Reason: The review contains figurative language that could confuse the model—e.g., "broken my heart" might be interpreted negatively, even though the context suggests a positive experience. Sentiment models are generally weak at detecting figurative expressions or idioms without additional semantic layers.