OPEN ACCESS

International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 2 (2025) https://doi.org/10.58723/ijaaiml.v2i2.403

Check for updates

Research Article

# A Deep Learning Approach to Sentiment Analysis of Hotel Reviews: Comparing BERT and LSTM Models

| Received: April 13, 2025 | Revised: May 22, 2025 | Accepted: June 14, 2025 | Publish: June 16, 2025 |
|---|---|---|---|

Gunawan Wang*, Mustafa Musa Jaber

**Abstract:**

**Background of study:** Background of study: The impact of online reviews on consumer behavior is especially relevant in the hospitality industry, and the sentiment corresponding to these reviews is difficult to determine due to the subjectivity involved in the reviews, disparate writing styles, and the noticeable class imbalance resulting from the positive reviews outnumbering the negative and neutral ones. Standard machine learning approaches are biased toward the majority class and do not address these problems well.

**Aims and scope of paper:** The present research uses BERT and LSTM deep learning models to perform classification of customer reviews for hotels into three categories: positive, neutral, and negative. The main focus of the research is to analyze the performance of the models concerning sentiment prediction and the handling of the data imbalance problem and to benchmark the models with and without the use of under-sampling.

**Methods:** The dataset comprising of 20,000 reviews from the TripAdvisor platform was pre processed in various ways including the removal of stop words/special characters, tokenization, stemming, and lemmatization. The customer reviews were assigned star ratings, which were aggregated into categories of 4-5 stars as positive, 3 stars as neutral, and 1-2 stars as negative. Random under-sampling was used to the positive class to achieve balance in the dataset. The BERT (bert-base-uncased) and LSTM models were prepared with what was assumed to be a final train-validation split of 80:20, and were evaluated based on standard metrics of accuracy, precision, recall, and rel F1 score, and with a cross-validation of 5 folds.

**Result:** Without the use of under-sampling, BERT achieved the best overall performance with an accuracy of 0.86 and an F1 score of 0.93 for the positive sentiment class and an F1 score of 0.79 in the negative sentiment class. However, all models struggled with neutral sentiments (BERT F1-score: 0.43, LSTM: 0.25). Under-sampling improved neutral class recall (BERT: 0.79) but decreased overall accuracy (BERT: 0.73; LSTM: 0.67) and positive class precision.

**Conclusion:** BERT generally outperforms LSTM for hotel review sentiment analysis, particularly with imbalanced data. While under-sampling helps address class imbalance by improving neutral recall, it incurs significant performance trade-offs, reducing overall accuracy and precision in majority classes due to information loss. Future work should explore advanced resampling (SMOTE, ADASYN) or transfer learning (RoBERTa, XLNet) for better balance and neutral sentiment classification.

Keywords: BERT, Class Imbalance, Hotel Reviews, LSTM, Sentiment Analysis.

## 1. INTRODUCTION

In the age of digitalization, online reviews have become a global phenomenon that is drastically changing the business landscape and consumer behavior (Verhoef et al., 2021). In the hospitality industry, these reviews are more than just feedback; they are invaluable digital assets, influencing a hotel's reputation, potential guests' booking decisions, and even operational strategies (Mishra et al., 2023). Modern consumers, accustomed to instant access to information, rely heavily on hotel reviews, which are usually rated using a star system, to gain insight into service quality and guest satisfaction before making a decision (Li et al., 2020). Platforms such as TripAdvisor, Booking.com and Google Reviews have become a major source of this information, allowing millions of travelers to share their experiences, both positive and negative (Xiang et al., 2017).

These reviews provide valuable feedback to businesses and help potential customers gauge the quality of a hotel or other services. However, extracting meaningful sentiment from these reviews is a complex and challenging task due to the subjective nature of human expression and the nuances involved in interpreting the text (Jim et al., 2024).

One of the main challenges in sentiment analysis of hotel reviews is the presence of an unbalanced data set,

where certain sentiment classes, such as positive reviews, are often overrepresented, while others, such as negative or neutral reviews, are underrepresented (Putra et al., 2025). This imbalance becomes more significant in the context of hotel reviews than in other domains due to several unique characteristics. First, self-selection bias is common; consumers tend to be more motivated to leave reviews when their experiences are either very good or very bad. This creates a "U-shaped" review pattern where positive reviews dominate, followed by extreme negative reviews, while neutral reviews or mediocre experiences are often ignored or under-reported (Hu et al., 2017).

An obstacle in the analysis of the hotels reviews sentiments is the unbalanced dataset reviews, where some sentiment classes, as in the case of the positive reviews, are often overrepresented, and some are underrepresented, such as the negative and neutral reviews (Putra et al., 2025).

The problem of imbalance in the dataset is therefore perpetuated. In essence, standard approaches, such as logistic regressions, decision trees, and support vector machines, are faced with problems arising from class imbalance, which leads to poorly defined predictions and biases. Despite the underperformance of the sentiment models across the different customer sentiment classes, especially towards the majority of the customers, they tend to perform poorly in capturing the sentiment for the minority sentiment categories (Hartmann et al., 2023).

This study will analyze the application of deep learning techniques and sophisticated natural language processing (NLP) methods to interpret the sentiments expressed in hotel reviews (Chi et al., 2025). The authors intend to examine potential utility star ratings from customer reviews to aid in the classification of the sentiment of the reviews. Reviews will convey sentiment in positive (4 and 5 stars), neutral (3 stars), and negative (1 and 2 stars) categories. This paper seeks the results of sentiment analysis from the classification of reviews using BERT (Bidirectional Encoder Representation from Transformers) and LSTM (Long Short-Term Memory) models (Tan et al., 2022). This paper aims to perform a comparative analysis of the two models in terms of their performance in the sentiment classification of reviews and sentiments of the hotel industry. Both models are highly regarded and considered superlative in the industry of NLP demonstrating high performance in multiple aspects of text classification. BERT is widely regarded and considered one of the best contextual models in text analysis in the industry and it is also a transformer model.

As noted, multiple linear models are primarily used to analyze reviews and text on a word-by-word basis (Supriyono et al., 2024). This model, however, is much more general and broad and is able to analyze text and reviews on a much greater level. This is especially the case for reviews with humor and sentiment, particularly cynical and negative sentiment, which are especially abundant in hotel reviews.

Adversarially, LSTM (Long Short-Term Memory), is a type of Recurrent Neural Network (RNN) that specializes in the formation of long-term dependencies in a given sequence. This specific feature of LSTM makes it the most suitable for sequence analysis, especially for reviews in the hotel industry, as it transcends the barriers of traditional models that fail to grasp the connection of distantly placed words in a given sentence (Mienye et al, 2024).

This study, besides contrasting these models, tackles the issue of class imbalance through undersampling of the sentiment class. Undersampling, by decreasing the sample size of the majority class, generates a balanced dataset that is believed to enhance machine learning models and reduce the bias toward the majority class (Mohammed et al, 2020). With the aim of elucidating which of the models performs the most accurate sentiment classification, we evaluated both models on a balanced dataset, and also assessed the effectiveness of undersampling to resolve the imbalance issue (George & Srividhya, 2022).

The most difficult aspect of the proposed research has already been discussed determining the sentiment of the reviews. This is difficult because of the subjectivity of customer reviews and the star rating imbalances. Consequently, this leads to the failure of conventional machine learning models. Thus, this research attempts to explain the potential of sophisticated deep learning models, such as BERT and LSTM, in the sentiment classification of an obscure and poorly structured hotel review system. The value of this research, compared to others, is the position and comparative study of BERT and LSTM architectures, which is one of the first most detailed and sophisticated works in deep learning and the hotel review sentiment classification problem. In this research, we intend to explain the impact of under-sampling on the two models, which is a gap we identified in the research concerning hotel review datasets.

Prior studies examined both BERT and LSTM separately or in different contexts; however, this study applies a different lens by focusing on which model better withstands data imbalance and under which conditions the accuracy of sentiment classification, particularly for the minority classes improves through under-sampling. We managed to construct insights for practitioners and academics concerning the model (s) to be utilized and the strategies to adopt in managing data imbalance to achieve a more refined and fair sentiment analysis of hotel reviews. This study attempts to demonstrate the positive effect of the most recent deep learning architectures on sentiment analysis and to offer practical recommendations for addressing the issue of class imbalance on review data sets.

## 2. MATERIAL AND METHOD

### Data Collection

The data acquired for this study comes from Laxel's (2020) TripAdvisor Hotel Reviews dataset available on Kaggle. The data set has 20,000 authentic user evaluations, each with a rating (1 to 5). This dataset is effective for sentiment analysis, text classification, and opinion mining in hospitality. The dataset contains two primary components, which are 'Reviews' (text) and 'Ratings' (integers). The dataset can be accessed at: https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

The dataset is highly successful at modeling user behavior, although the rating imbalance skews positively, especially when rating anger is sentiment classified. The rating sentiment imbalance is most pronounced if the dataset is used for positive/negative sentiment classification. Prior to sentiment classification, the initial set of data demonstrates the bias in user generated sentiment, which is the primary focus for this dataset. The initial bias sentiment for this data is, prior to sampling, as follows:

1. Positive Reviews (4 and 5 stars): [Enter the exact number/percentage here, for example, 15,000 reviews/ 75%].
2. Neutral Reviews (3 stars): [Enter the exact number/percentage here, for example, 2,000 reviews/ 10%].
3. Negative Reviews (1 and 2 stars): [Enter the exact number/percentage here, for example, 3,000 reviews/ 15%] (Roumeliotis et al., 2024).

The first analysis shows the data distribution, which reflects a positive rating bias. Class imbalance occurs when there is a difference in the distribution of class labels (in this case, positive reviews). This data can be used for developing understanding of customer satisfaction, features prioritization in the reviews, and customer satisfaction prediction systems, among other things. Preprocessing is particularly useful for systems that analyze customer feedback and improve services in the hospitality sector, although it lacks valuable metadata, such as the hotel name, user information, and review dates.

### Data Pre-processing

The reviews in the dataset are accompanied by star ratings on a scale of 1 to 5. For the purpose of the sentiment analysis, the star ratings are transformed to three classes of sentiments. These are positive for ratings of 4 and 5 stars, neutral for 3 stars, and negative for 1 and 2 stars (Arroni et al. 2023). During the review preprocessing, stop words, special characters, and tokens are removed. Then, text normalization is carried out through stemming and lemmatization. Once this is done, the text data will be ready to be analyzed and input to the sentiment classifiers.

### Handling Imbalanced Data

The data set shows significant imbalance which is mostly skewed in favor of positive reviews due to the data collection methodology stated in the data collection section. Such imbalance has the potential to severely hinder the performance of the machine learning models by accommodating more bias to the majority class and generalizing poorly to the minority class. To resolve this, and to avoid the model being too biased, random under-sampling was done to the majority class (positive reviews) in order to even out the sentiment label distribution (Chamidah et al, 2024).

Random under-sampling was opted for due to its ease of use and efficiency with shrinking the majority class so that a more balanced data set can be achieved. Even though techniques like over-sampling and hybrid techniques can be used, random under-sampling is still the most effective for this particular case since the majority class is very large and significant reduction can be achieved without losing too much information (Miftahushudur et al, 2025). This technique enables the model to fairly represent each of the sentiment classes during the training and as a result, more accurate and confident classification across the different classes.

### Model Selection

For the purpose of conducting sentiment analyses, two deep learning models were selected owing to their success regarding text-based tasks and sophisticated natural language processing capabilities. One of the models chosen for the study is BERT (Bidirectional Encoder Representations from Transformers). We chose BERT because it is a pre-trained transformer model, meaning it will be able to analyze the relationships and understand the context of a given phrase by analyzing the words in different directions (Bidirectional), and is thus better equipped to understand the nuances of specific phrases in a given review (Gardazi et al., 2025). Among the different BERT models, we decided to use the model called bert-base-uncased, which has 12 layers, 768 hidden units, and 12 attention heads. For the second model, we chose an LSTM (Long and Short Term Memory) model, which in this case, possesses the capability of capturing long-range dependencies in sequential data. This is particularly pertinent to sentiment analysis, because specific words throughout a review can insinuate different feelings (Malashin et al., 2024). The LSTM model for this study is comprised of two LSTM layers and each layer has 128 units, and there is a dense final output layer. A pre-trained GloVe embedding layer containing 100-dimensional vectors was used for the word embeddings.

Overall, given the advancements in the field, it is expected that the LSTM and BERT models will analyze the text thoroughly and perform the task excellently.

### Training and Evaluation

The same pre-processed dataset used to train the BERT and LSTM models has been divided into training and validation splits. The training splits assist model

development, while the validation splits serve to fine-tune model hyperparameters and assess performance. The models are evaluated for the consistency and correctness of their sentiment classification, using a wide array of assessment metrics: overall accuracy, precision, recall, and the F1 score. To assess the models and mitigate overfitting, the models are subject to cross-validation. This step is primarily aimed at serving a comprehensive assessment of the models' performance from multiple vantage points.

## 3. RESULT AND DISCUSSION

### 3.1 Results

The performance of BERT and LSTM models on sentiment analysis of hotel reviews, further, no under-sampling to mitigate class imbalance was performed.

Both models were assessed based on different metrics, which include precision, recall, F1-score, and overall accuracy. The primary task was to classify the reviews into three different sentiments: positive (4 and 5 stars), neutral (3 stars), and negative (1 and 2 stars). Even so, both the BERT and LSTM models were able to demonstrate different degrees of detecting and classifying each of the sentiments in the hotel reviews. Each of the models differentiated themselves in all of the given sentiments as will be elaborated on in subsequent sections. The comparative results are shown in Table 1.

#### Comparison of Model Performance

In Table 1, we see the thorough comparison of the performance of the two deep learning models BERT and LSTM in the task of multi-class sentiment classification. Each model underwent an evaluation with and without under-sampling to mitigate possible class imbalance, while accuracy, precision, recall, and F1-score were selected as the performance evaluation metrics.

**Table 1.** The comparison results

| Approaches | Accuracy | Class | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| BERT Model | 0.86 | Negative | 0.80 | 0.79 | 0.79 |
| | | Neutral | 0.41 | 0.45 | 0.43 |
| | | Positive | 0.94 | 0.93 | 0.93 |
| BERT Model (w/ under sampling) | 0.73 | Negative | 0.86 | 0.61 | 0.72 |
| | | Neutral | 0.26 | 0.79 | 0.39 |
| | | Positive | 0.97 | 0.75 | 0.85 |
| LSTM Model | 0.84 | Negative | 0.63 | 0.83 | 0.72 |
| | | Neutral | 0.42 | 0.17 | 0.25 |
| | | Positive | 0.92 | 0.94 | 0.93 |
| LSTM Model (w/ under sampling) | 0.67 | Negative | 0.64 | 0.64 | 0.64 |
| | | Neutral | 0.20 | 0.59 | 0.29 |
| | | Positive | 0.96 | 0.69 | 0.80 |

BERT model (without undersampling) achieved the highest accuracy at 0.86 for all models tested. This is the best result in this scenario. Examining class F1 scores, positive (0.93) and negative (0.79) classes, without undersampling BERT, performed well. This demonstrates the ability of this model to identify the respective sentiments. 0.43 does represent a substantial drop in F1 score to the neutral class, where the model only achieved 0.43. One can assume that in this case it might be due to the neutral subclass having less frequent and therefore, possibly result in class imbalance to other sentiment classes.On the other hand, the LSTM model, without under-sampling, achieved an overall accuracy of 0.84, which is still lower than BERT. It also succeeded in the classification of positive sentiments with an F1 score of 0.93, however, compared to BERT, LSTM performed worse in the neutral class with an F1 score of 0.25, which was the result of a low recall of 0.17. The negative class was moderately performed with an F1 score of 0.72. This shows that the LSTM model is positive and negative sentiments and is less successful with the neutral class and other less infrequent classes. Therefore, it indicates that in the combination of neutral

and negative sentiments, BERT has a stronger contextual understanding in comparison to LSTM.

### Under-sampling Impact

Applying under-sampling to BERT caused its accuracy to fall to 0.73 from 0.86. The F1-scores for the positive and negative classes decreased to 0.85 and 0.72, respectively. More recall for the neutral class from 0.45 to 0.79, however, came at the low precision of 0.26, meaning that the neutral F1-score only marginally increased to 0.39 and only a little to 0.79 the neutral recall. This indicates under-sampling improved BERT's ability to identify neutral instances and, consequently, fewer neutral cases were misclassified as non-neutral. More neutral cases were classified as neutral, though, meaning more instances which were non-neutral were incorrectly classified as neutral.

The LSTM model subsequently demonstrated even lower accuracy with under-sampling at 0.67, the worst score of any model. The neutral class F1 score did improve slightly to 0.29, but the declines in negative and positive class F1 scores more than countered this. The F1 score in the positive class dropped from a robust 0.93 to 0.80, and the negative class F1 score also dropped to 0.64. The LSTM model's overall accuracy and ability to classify dominant positive and negative classes is negatively impacted by under-sampling and confirms that increasing recall for the negative class comes at a cost in overall accuracy and precision for the positive dominant class.

### Comparison Metric

Figure 1 graphically presents how performance from 4 models during the sentiment analysis of hotel reviews compares against each model performance metric of precision, recall, F1 score, and accuracy. Each model performance metric is a function of how well that model categorizes hotel reviews into positive, negative, and neutral sentiment. The performance metrics and models illustrate the extent of class imbalance, in addition to the inherent difficulty of analyzing customer reviews.

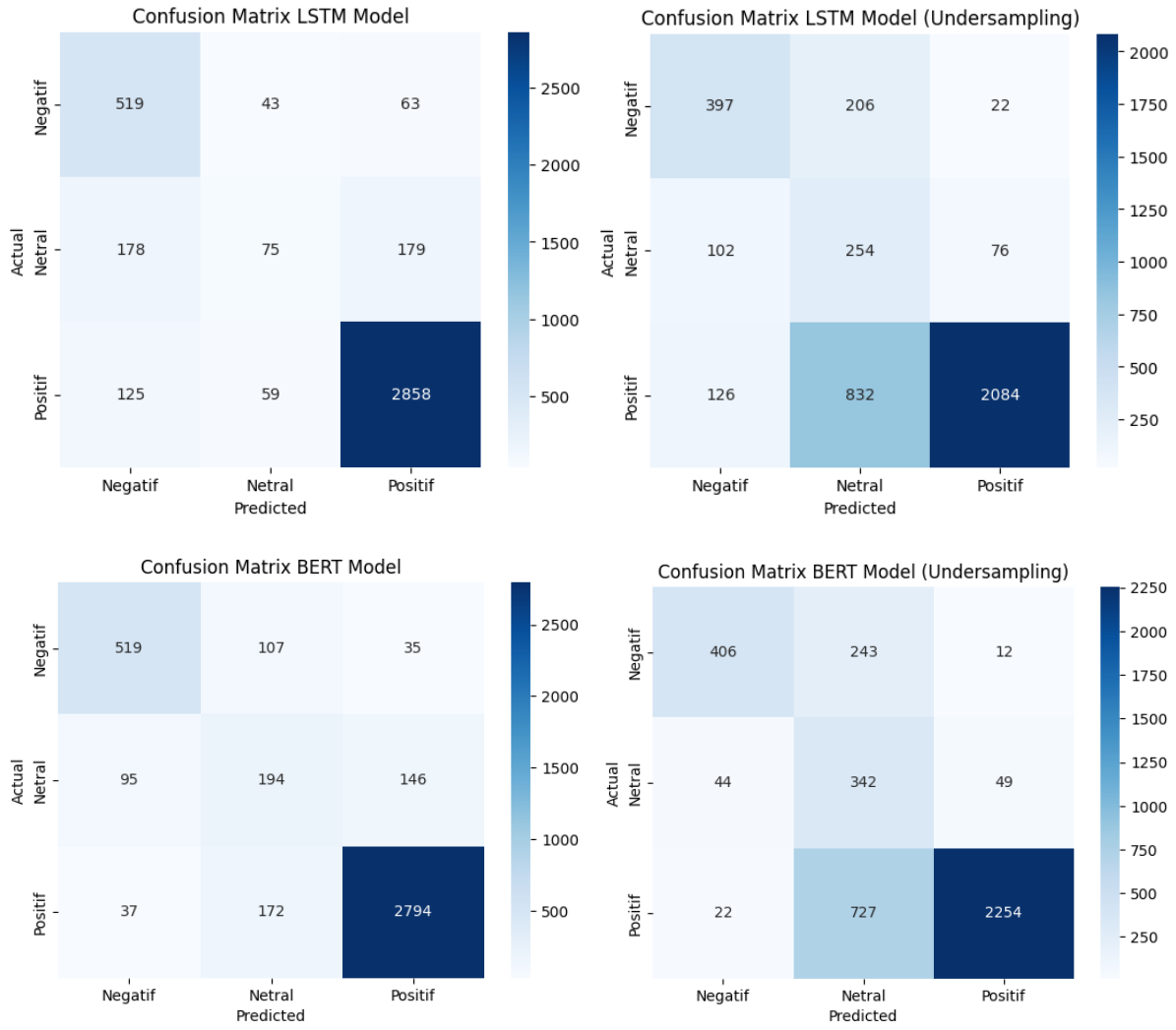


**Figure 1.** The confusion matrices for each different model.

The confusion matrices in Figure 1 allow for a more nuanced understanding of the behavior of each model BERT and LSTM with and without undersampling. For the LSTM model without under-sampling, there appears to be a predominant bias toward the Positive class, likely due to it being positively predicted more often, which may bias the model. The model attained a high Positive predictive count, yet the Neutral examples were predicted incorrectly, and a majority of them were mislabeled to either the Negative or Positive class. This partially illustrates the inability of the LSTM model to identify more nuanced and differentiated granular sentiments and the effect of data imbalance particularly in the negative and neutral sentiments.

Considering under-sampling for the LSTM, the prediction distribution for the Neutral class improves, given the higher number of correctly classified Neutral class instances. Under-sampling in the LSTM has achieved an improved distribution for Neutral, but it comes at the expense of increased Neutral/Positive misclassifications and reduced overall correct classifications of Positive sentiments. Notably, this depicts the typical effect of under-sampling, where the majority class of training examples is reduced, which results in diminished overall generalization.

The current setup shows that BERT with no under-sampling is overall better than LSTM. It has a fair predictive ability for Positive sentiments and is marginally better than LSTM for Negative and Neutral classes. However, Neutral is still a problematic class, with significant misclassification. Even with the BERT model's robustness, this shows the impact of class imbalance.

After under-sampling was performed on BERT, the model shows a marked improvement for Balanced Accuracies across the three classes of sentiments. Particularly, the Neutral class shows a better predictive ability classification. Even though the Positive predictive classification shows a slight decline, a more equitable distribution of classifications has been achieved, which is a beneficial trade-off for a number of practical applications where classification across all classes is necessary and demands equal importance.

### 3.2 Discussion

The goal of this study was to evaluate the use of under-sampling as a means of tackling class imbalance in the Negative, Neutral, and Positive multi-class sentiment classification problem. It is likely that the original distributions of the data contained a greater number of instances of the Positive class, which led to the models focusing on this majority class while ignoring the less represented Negative and Neutral classes.

To address class imbalance, the majority class was reduced using under-sampling techniques. This was to allow the model to train on all classes equally and improve detection and positive classification of the underrepresented class of the Neutral and Negative sentiments. This is useful in scenarios where capturing the minority class is more critical than overall accuracy.

The results, however, showed that even though under-sampling improved Neutral class recall (worst in BERT model) a negative accuracy trade off was present. This resulted in a loss of recall, precision, and F1 of Positive class and overall model accuracy, and even higher Optimized Positive class F1 scores in the initial models. This was a result of the majority class being cut out and a loss of the models ability to learn sufficient class patterns.

Thus, although the negative effects of class imbalance were partially mitigated through the use of under-sampling, important information was lost and the overall efficacy of the model was compromised. This exemplifies the drawbacks of under-sampling and strengthens the case for considering other methods, such as class weighting, oversampling, or even the creation of synthetic data, as in the case of SMOTE, which may enhance performance in the under-represented class while maintaining the model's generalization capacity.

#### 3.2.1 Implications

The current research offers valuable contributions to the field of sentiment analysis, particularly for the case of hotel reviews, for both practitioners and researchers. It shows that BERT and other sophisticated deep learning models outperform all earlier techniques in sentiment analysis. This matters greatly to the hospitality sector because the accurate sentiment analysis of online reviews sales, reputation, and operational planning. The study highlights the challenges of categorizing sentiment as neutral, which suggests further complexity in the description of the vague and sentimentally deficient. While undersampling attains greater recall for the neutral minority class, it does so at the expense of precision and overall accuracy of the dominant class. This suggests that the strategy used for bearing class imbalance in data sets ought to place emphasis on the applied objectives, especially in circumstances when the goal is to achieve substantial recall in the minority class at the expense of precision.

#### 3.2.2 Research contribution

The current study stands out in the field literature for having conducted the first extensive comparison between LSTM and BERT models for the task of sentiment classification on unbalanced hotel reviews. It explicitly points out the model that exhibits the greatest robustness to data imbalance and the model most under-sampled for negative and neutral sentiment classifications. It provides recommendations on model choice and data imbalance mitigation techniques to improve the accuracy and bias of sentiment analyses on hotel reviews. Additionally, it consolidates the evidence that BERT and similar transformer models remain the most suitable for advanced sentiment analysis tasks, owing to their capacity to capture and reason with deeper levels of semantics, while also noting that the

classification of neutral sentiment is an important area that requires further development.

### 3.2.3 Limitations

A significant limitation for all models was the inability to truly identify neutral sentiments. This is especially true given recent deep learning advancements with models like BERT and LSTM. The neutral language is especially tricky, and it compounds the problem to know that reviews classifying neutrally are less frequent. Class imbalance was therefore the main issue. Though the study used under-sampling to attempt to correct for class imbalance, the study demonstrated that it led to a significant trade-off. This was the reduction overall accuracy, as well as reduction in precision and F1-score in the majority class (positive). This means that under-sampling causes a loss for the majority class from which there is typically plenty of data, therefore the overall generalization to the model is decreased. The dataset is also missing pieces like the hotel name, the user, demographics, timestamps, etc. All of which are required for a more in depth analysis.

### 3.2.4 Suggestions

In future work, there are a number of strategies that can be used to improve the analysis of sentiment in hotel reviews. The first and most simple is the use of more advanced resampling techniques, which can be applied in place of random under-sampling. For example, SMOTE (Synthetic minority oversampling technique) or ADASYN can be used to create synthetic samples for the underrepresented classes of sentiment data, and especially for neutral sentiment reviews.

This method attempts to offer the model a perspective that does not compromise information loss from minimizing the majority class. An alternative method that may present success is using transfer learning with other pre-trained models like RoBERTa, DistilBERT, or XLNet. These models may encompass the range of subtle meanings that can exist in hotel reviews, leading to the improved classification of sentiments that may be ambiguous. Building on the other approaches seeks to construct a more advanced sentiment analysis tool that is capable of recognizing class imbalance and contextual nuances of hotel reviews.

## 4. CONCLUSION

This study performs an analysis of BERT and LSTM architectures for sentiment analysis of multi-class hotel reviews (Negative, Neutral, Positive), incorporating the effects of under-sampling to mitigate class imbalance. The unique aspect of this study is the comprehensive comparative assessment of the two prominent deep learning frameworks in the domain of unbalanced sentiment data, and more specifically, the difficulties encountered in classifying neutral sentiment.The results indicate that the BERT model without under-sampling has a best overall performance, and also the highest accuracy of 0.86. The model demonstrated a strong capability of identifying both positive (F1-score: 0.93) and negative (F1-score: 0.79) reviews. Yet, it appears that both BERT and LSTM models had a consistent tendency to misclassify (or fail to classify) neutral reviews, and had lower F1-scores (BERT: 0.43; LSTM: 0.25 without under-sampling). This could indicate the nuanced difficulty of neutral sentiment and the class imbalance of the left-out minority class.

The use of under-sampling, had as its purpose mitigating the class imbalance, but the result is of a notable performance trade-off. While under-sampling resulted in a slight improvement in recall for the neutral class (most notably for BERT, from 0.45 to 0.79) it resulted in a decrease in overall accuracy (BERT from 0.86 to 0.73; LSTM from 0.84 to 0.67), and a significant decrease in precision and F1-score, in the (positive) majority class. This suggests that the loss of data from the majority class can result in the loss of useful data, and can diminish the model's overall ability to generalize.

The most noteworthy conclusion drawn from this research is that, among various sentiment analysis systems, transformer-based architectures like BERT, due to their capability to grasp deeper levels of semantic intricacies, outperform all others, including LSTMs. Despite this, the difficulty of classifying neutral sentiment is a substantial problem that remains. While under-sampling is a strategy that assists in the mitigation of imbalance, the adverse effects that accompany its use are substantial, and thus more advanced methods to tackle class imbalance are warranted.

To improve the sensitivity analysis of hotel reviews, one of the strategies is Advanced Resampling Methods, while the other is Transfer Learning with Other Pre-trained Models. Models such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN can create new synthetic data for the neutral reviews, which are the underrepresented classes, so that the model can see the data from a balanced perspective without losing the majority class information. Moreover, other pre-trained models, such as RoBERTa or DistilBERT, or even XLNet, used in transfer learning, can improve the sensitivity analysis of the model even further, as they may capture nuanced semantics of the hotel reviews more effectively. With the integration of these approaches, we can create a more robust and advanced system for the analysis of the sensitivity of hotel reviews, and even more advanced class imbalance, in addition to the nuanced and contextual understanding of hotel reviews.

## 5. ACKNOWLEDGEMENT

This research was supported by Bina Nusantara University, Jakarta, and Dijlah University College, Iraq. The authors also wish to thank the reviewers for their insightful comments and suggestions that helped improve the quality of this manuscript.

## 6. AUTHOR CONTRIBUTION STATEMENT

GW: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing Original Draft, Visualization, Project Administration. MMJ: Supervision, Writing Review & Editing.

## AUTHOR INFORMATION

### Corresponding Authors

Gunawan Wang, Bina Nusantara University, Jakarta
https://orcid.org/0000-0002-0877-9966
Email: gunawan.wang@binus.ac.id

### Authors

Mustafa Musa Jaber, Computer Sciences Department, Dijlah University College, Iraq
https://orcid.org/0000-0002-4416-2162
Email: Mustafa.jaber@turath.edu.iq

## REFFERENCE

Arroni, S., Galán, Y., Guzmán-Guzmán, X., Núñez-Valdez, E. R., & Gómez, A. (2023). Sentiment Analysis and Classification of Hotel Opinions in Twitter With the Transformer Architecture. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(1), 53–63. https://doi.org/10.9781/ijimai.2023.02.005

Chamidah, N., Widiyanto, D., Seta, H. B., & Aziz, A. A. (2024). The Impact of Oversampling and Undersampling on Aspect-Based Sentiment Analysis of Indramayu Tourism Using Logistic Regression. *Revue d'Intelligence Artificielle*, *38*(3), 795–804. https://doi.org/10.18280/ria.380306

Chi, D., Huang, T., Jia, Z., & Zhang, S. (2025). Research on sentiment analysis of hotel review text based on BERT-TCN-BiLSTM-attention model. *Array*, *25*(February), 100378. https://doi.org/10.1016/j.array.2025.100378

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: a review. *Artificial Intelligence Review*, *58*(6). https://doi.org/10.1007/s10462-025-11162-5

George, S., & Srividhya, V. (2022). Performance Evaluation of Sentiment Analysis on Balanced and Imbalanced Dataset Using Ensemble Approach. *Indian Journal of Science and Technology*, *15*(17), 790–797. https://doi.org/10.17485/ijst/v15i17.2339

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hu, N., Pavlou, P. A., Zhang, J., Hu, N. ;, & Pavlou, P. A. ; (2017). On self-selection biases in online product reviews On self-selection biases in online product reviews Part of the Databases and Information Systems Commons, E-Commerce Commons, and the Numerical Analysis and Scientific Computing Commons Citation Citation . *MIS Quarterly*, *41*(2), 449–472. https://doi.org/10.25300/MISQ/2017/41.2.06

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, *6*(February), 100059. https://doi.org/10.1016/j.nlp.2024.100059

Li, H., Liu, Y., Tan, C. W., & Hu, F. (2020). Comprehending customer satisfaction with hotels: Data analysis of consumer-generated reviews. *International Journal of Contemporary Hospitality Management*, *32*(5), 1713–1735. https://doi.org/10.1108/IJCHM-06-2019-0581

Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2024). Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review. *Polymers*, *16*(18), 1–44. https://doi.org/10.3390/polym16182607

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

Miftahushudur, T., Sahin, H. M., Grieve, B., & Yin, H. (2025). A Survey of Methods for Addressing Imbalance Data Problems in Agriculture Applications. *Remote Sensing*, *17*(3), 1–31. https://doi.org/10.3390/rs17030454

Mishra, A., Kishan, K., & Tewari, V. (2023). THE INFLUENCE OF ONLINE REVIEWS ON CONSUMER DECISION-MAKING IN THE HOTEL INDUSTRY. *Journal of Data Acquisition and Processing*, *3*(September), 2559–2573. https://doi.org/10.28934/jwee23.34.pp48-74

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, *May*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Putra, P. P., Anam, M. K., Chan, A. S., Hadi, A., Hendri, N., & Masnur, A. (2025). Optimizing Sentiment Analysis on Imbalanced Hotel Review Data Using SMOTE and Ensemble Machine Learning Techniques. *Journal of Applied Data Sciences*,

*6*(2), 936–951. https://doi.org/10.47738/jads.v6i2.618

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Leveraging Large Language Models in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification and Sentiment Analysis. *Information (Switzerland)*, *15*(12), 1–23. https://doi.org/10.3390/info15120792

Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, *16*(April), 100173. https://doi.org/10.1016/j.teler.2024.100173

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access*, *10*, 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, *122*(November 2019), 889–901. https://doi.org/10.1016/j.jbusres.2019.09.022

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*(February), 51–65. https://doi.org/10.1016/j.tourman.2016.10.001