Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

Improved Sentiment Classification Using Multilingual BERT with Enhanced Performance Evaluation for Hotel Guest Review Analysis

Yerik Afrianto Singgalen

Faculty of Business Administration and Communication, Tourism Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: yerik.afrianto@atmajaya.ac.id
Submitted: 29/01/2025; Accepted: 28/02/2025; Published: 28/02/2025

Abstract−Sentiment analysis in hotel guest reviews has become essential for evaluating customer satisfaction and service quality. This study improves sentiment classification accuracy by utilizing the Multilingual BERT model with an improved performance evaluation framework. Using the Knowledge Discovery in Databases (KDD) methodology, this research involves data selection, preprocessing, transformation, sentiment classification, and performance evaluation. A dataset of 715 hotel reviews from Qubika Boutique Hotel, sourced from Agoda, was used to assess the model's effectiveness. The classification results showed high accuracy in identifying positive sentiment, with 98% precision, 97% memory, and 98% F1 score, as observed in 432 correctly classified reviews. However, challenges were identified in the classification of neutral sentiment, which achieved a precision of 87% with 127 correctly classified cases, and negative sentiment, where the accuracy was 92%, with 104 correctly identified reviews. The overlap in confidence scores, especially in the range of 0.4-0.6 between neutral and negative sentiment, highlights the need for improved contextual embedding and hybrid modeling techniques. The sentiment distribution analysis revealed that 60-70% of reviews were positive, 20-30% neutral, and 10-15% indicated dissatisfaction, underscoring the need for targeted service improvement. These findings provide valuable insights for data-driven decision-making in hospitality management, enabling businesses to strengthen service power and address critical areas of concern. Future research should focus on refining model interpretability, expanding multilingual datasets, and integrating real-time sentiment analysis to improve classification performance. Strengthening these aspects will contribute to a more robust and scalable sentiment analysis framework, ensuring greater precision in capturing the guest experience and optimizing service strategies in the hospitality industry.

Keywords: Sentiment Analysis; Multilingual BERT; Hospitality Management; Performance Evaluation; Data-Driven Decision-Making

1. INTRODUCTION

Sentiment analysis in hotel guest reviews has become essential in computational linguistics and artificial intelligence, especially integrating deep learning architectures such as Multilingual BERT. Sentiment classification systems require robust methodologies to accommodate linguistic variations and contextual intricacies in multilingual datasets, ensuring higher accuracy and generalization [1], [2]. Conventional approaches often present limitations in cross-language comprehension and performance evaluation, which necessitates improvements in the adaptability and interpretability of the model [3]–[5]. Combining advanced performance evaluation metrics and refinement techniques strengthens sentiment classification models by reducing bias, improving contextual embedding, and optimizing classification results. Enhancements to sentiment analysis through enhanced multilingual embedding advance machine learning applications in the hospitality sector. This fosters more precise consumer sentiment insights, strengthening the data-driven decision-making process in service management.

The urgency of this research is underscored by the increasing reliance on sentiment classification models to extract meaningful insights from large amounts of unstructured textual data in the hospitality sector. The rapid expansion of online guest reviews has required a more sophisticated analytical framework to distinguish nuanced sentiments across various linguistic and cultural contexts [6], [7]. Existing sentiment analysis methodologies often face challenges in handling multilingual data, contextual ambiguity, and performance inconsistencies, leading to suboptimal classification results [8]–[10]. Integrating enhanced multilingual models and enhanced evaluation metrics is essential to improve the accuracy and reliability of sentiment classification systems. Strengthening these analytical capabilities facilitates a more accurate interpretation of consumer sentiment and empowers data-driven strategies that optimize service quality, customer engagement, and business competitiveness in the hospitality industry.

This study aims to improve sentiment classification accuracy in hotel guest reviews by utilizing a multilingual BERT model with optimized performance evaluation techniques. The complexity of sentiment analysis in the hospitality sector arises from the diverse linguistic expressions, contextual variations, and implicit sentiments embedded in textual data [11]. Conventional sentiment classification models often struggle with these challenges, resulting in inconsistencies and reduced interpretability in cross-language applications. By integrating advanced model refinement strategies and comprehensive evaluation metrics, this study aims to refine the sentiment classification methodology, ensuring higher reliability and adaptability in a multilingual environment [12]–[14]. Strengthening analytical precision in sentiment classification contributes to an adequate

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

interpretation of consumer sentiment, facilitating data-driven decision-making processes that improve service quality and customer experience management in the hospitality industry.

The Knowledge Discovery in Database (KDD) methodology was used in this study to extract valuable insights from a systematic review of hotel guests through a structured data mining process. This approach includes several critical stages, including data selection, preprocessing, transformation, data mining, and interpretation, ensuring a comprehensive analytical framework for sentiment classification [15]–[17]. The complexity of multilingual sentiment analysis requires a careful data handling process, where noise reduction, feature extraction, and model optimization play a crucial role in improving classification accuracy [18], [19]. By utilizing KDD, sentiment classification benefits from a structured and iterative refinement process, allowing for the identification of complex sentiment patterns across various linguistic contexts. Applying this methodology improves analytical precision, strengthens the reliability of sentiment classification models in the hospitality sector, and facilitates more effective data-driven decision-making.

Existing studies on sentiment classification in hotel guest reviews have primarily focused on conventional machine learning techniques and deep learning models, showing significant advances in text analysis and opinion mining. Various approaches have improved sentiment classification accuracy across languages and contexts, including Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), and Transformer-based architectures [20]–[22]. However, challenges remain in handling multilingual datasets, optimizing model interpretability, and refining performance evaluation metrics, often resulting in inconsistencies in sentiment predictions [23]–[26]. The research gap lies in the need for a more adaptive and linguistically diverse sentiment classification framework that integrates multilingual BERT with robust performance evaluation mechanisms. Overcoming these limitations contributes to developing more precise and measurable sentiment analysis models and improving decision-making processes in the hospitality industry through better interpretation of customer sentiment.

The novelty of this study lies in integrating multilingual BERT with an advanced performance evaluation framework to improve the accuracy of sentiment classification in hotel guest reviews. Existing sentiment analysis models often face limitations in handling cross-language variation, contextual ambiguity, and unbalanced data sets, which affect the reliability of the classification [27]. By refining the model's adaptability through optimized tuning techniques and incorporating comprehensive evaluation metrics, this study introduces a more robust approach to multilingual sentiment classification [28]. The proposed framework improves predictive precision, interpretability, and scalability across linguistic datasets. Strengthening sentiment analysis methodologies in this way contributes to more reliable consumer sentiment insights, supporting strategic decision-making processes in the hospitality industry.

The theoretical contribution of this study is reflected in the advancement of sentiment classification methodology through integrating multilingual BERT with an improved performance evaluation framework. Refining machine learning techniques in processing multilingual textual data expands existing knowledge in computational linguistics and deep learning applications, especially in sentiment analysis in the hospitality sector [29]–[31]. From a practical perspective, implementing this model provides a more accurate and adaptable sentiment classification system, allowing businesses to extract meaningful insights from diverse guest reviews with better reliability [32]–[34]. This enhanced analytical approach supports a more precise interpretation of consumer sentiment, facilitating strategic decision-making in service management, customer engagement, and market positioning. Strengthening the theoretical foundation and practical implementation in this domain contributes to developing a sustainable, intelligent sentiment analysis system and strengthening data-driven innovation in the hospitality industry.

Future research should explore refinements of sentiment classification models by incorporating more sophisticated deep learning architectures and expanding the scope of multilingual datasets. The increasing complexity of sentiment analysis, especially in the hospitality sector, requires continuous improvements in model interpretability, contextual understanding, and computational efficiency [35]. Integrating transformer-based models with hybrid approaches, such as knowledge graphs or domain-specific embeddings, can improve sentiment classification accuracy while addressing semantic nuances in diverse textual data. Additionally, applying sentiment analysis in a real-time environment and its integration with predictive analytics will provide valuable insights for dynamic decision-making. Strengthening these aspects in subsequent studies will contribute to developing more sophisticated, adaptive, and practical sentiment analysis frameworks, supporting more effective data-driven strategies across various industries.

2. RESEARCH METHODOLOGY

2.1 Research Workflow

The research workflow adheres to the Knowledge Discovery in Database (KDD) framework, systematically guiding the process from data collection to interpretation [36]. The process begins with data acquisition, which involves identifying relevant sources and extracting a structured dataset for analysis. After this, data preprocessing techniques are used, which include noise reduction, handling of missing values, and feature

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

validation to improve data quality [37]. Furthermore, data transformation integrates contextual embedding through BERT-based tokenization, optimizing inputs for model training. The sentiment classification phase uses advanced deep learning algorithms to ensure accurate predictions, supported by rigorous performance evaluation metrics such as confusion matrices and ROC-AUC curves. The workflow concludes with a knowledge presentation, including predictive reporting and insightful summaries, establishing a structured and efficient approach to sentiment analysis in a multilingual context.

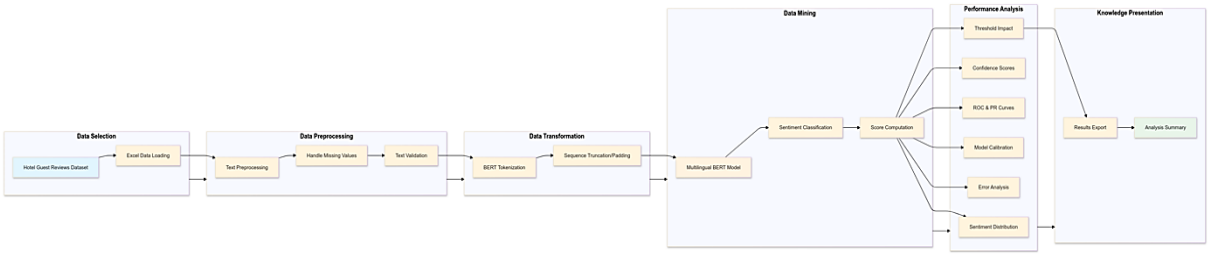

Figure 1. Research Workflow

Figure 1 illustrates a comprehensive research workflow aligned with the Knowledge Discovery in Databases (KDD) framework to ensure a systematic and efficient process. The workflow begins with the data acquisition stage, which involves identifying relevant data sources and extracting data sets essential for sentiment classification. The following preprocessing phase focuses on refining the dataset through noise reduction, handling missing values, and feature validation to improve the analytics quality. After this, the data transformation is performed using advanced BERT-based tokenization to prepare the data for classification. The sentiment classification stage uses machine learning models to analyze and classify sentiment, supported by rigorous evaluation metrics such as confusion matrices and performance score curves. The final stage involves presenting knowledge, synthesizing the results into predictive reports and analytical summaries, and providing actionable insights. This structured workflow demonstrates a robust approach to managing and analyzing data, ensuring the reliability and applicability of findings.

Figure 2. Data Selection, Preprocessing, and Transformation

Figure 2 illustrates the sequential data selection stages of preprocessing, preprocessing, and transformation, emphasizing the basic steps for effective sentiment classification. The process begins with data selection, which entails identifying relevant guest review datasets and extracting them for further analysis. Furthermore, the preprocessing phase ensures data quality by addressing missing values, reducing noise, and validating textual data to improve analytics resilience. After preprocessing, the transformation stage uses advanced BERT-based tokenization to transform raw textual data into contextual embeddings optimized for machine learning models. This structured approach ensures the data is prepared systematically, allowing for a more accurate and reliable sentiment classification process. These stages collectively form a critical path that strengthens the analytical precision and application of findings in multilingual sentiment analysis.

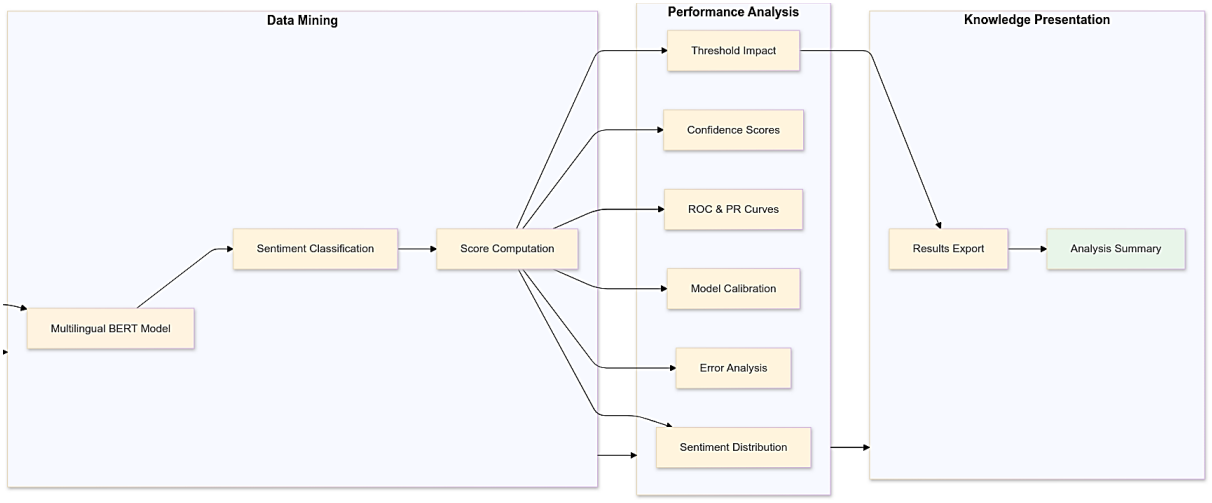

Figure 3. Performance Analysis and Knowledge Presentation

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

Figure 3 illustrates the stages of performance analysis and knowledge presentation, emphasizing the importance of evaluating and communicating sentiment classification results. This process begins with sentiment classification and score calculation, where the multilingual BERT model generates predictions and assigns confidence scores to classify sentiment. The performance analysis thoroughly examines the impact of thresholds, ROC and PR curves, model calibration, error distribution, and sentiment distribution, ensuring a comprehensive understanding of the model's strengths and limitations. This stage provides insight into the reliability and accuracy of the classification results. The final stage, knowledge presentation, involves exporting results and generating a summary of the analysis, turning raw data into actionable insights. This structured process allows stakeholders to make decisions based on robust analytical findings, ensuring the relevance and usefulness of sentiment analysis in practical applications.

2.2 Datasets

The dataset used in this study consists of 715 rows of review data collected from Qubika Boutique Hotel through the Agoda platform. This dataset is a vital resource, capturing diverse guest experiences and satisfaction levels and allowing for a comprehensive analysis of sentiment trends. Including real-world reviews ensures the applicability and relevance of the findings, providing valuable insights into customer perceptions and preferences. Leveraging this dataset, the study addresses the need for a data-driven approach to improve service quality and align with evolving guest expectations. This robust dataset not only facilitates a detailed exploration of sentiment classifications but also reinforces the importance of empirical evidence in advancing the analytical capabilities of the hospitality sector.

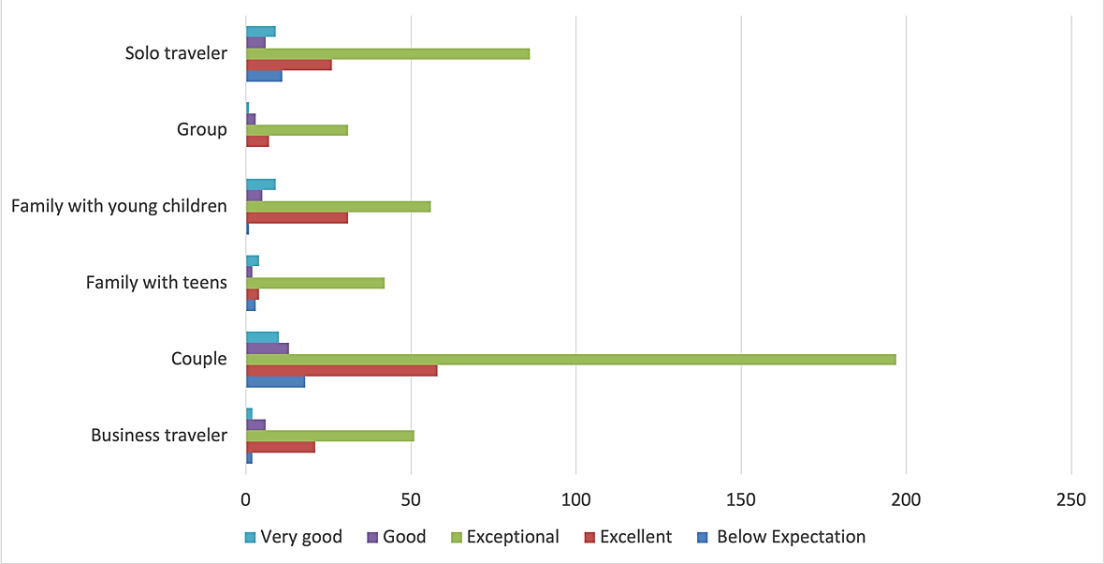

Figure 4. Expectations by Visitor Type (715 Reviews)

Figure 4 illustrates the distribution of visitor expectations across different types of travelers based on 715 reviews, highlighting different satisfaction levels. Couples make up the majority, with a significant preference for "Good" and "Very Good" ratings, showing moderate to high satisfaction among this group. In contrast, business and solo travelers showed a more balanced distribution across categories, with the notable example of "Below Expectations," which showed diverse experiences within these groups. Families with young children and teens prefer higher ratings such as "Excellent" and "Excellent," reflecting a tendency toward more favorable experiences. The data shows that traveler type significantly influences expectations and satisfaction levels, underscoring the need for tailored hospitality strategies. Understanding these variations provides valuable insights to improve the quality of service and visitor experience across different demographic groups.

The distribution of satisfaction levels across different types of travelers reveals significant variations in experiences and expectations. Couples dominate the number of reviews, with the majority rating "Good" or "Very Good," accounting for about 40-50% of the total, showing consistently high levels of satisfaction within this group. In contrast, business and solo travelers displayed a more even spread across all categories, including 15–20% who stood out in "Below Expectations," highlighting a broader range of experiences and challenges in meeting their diverse expectations. Families with young children and teens show a higher tendency towards "Excellent" and "Excellent" ratings, with 30-40% of reviews reflecting a favorable experience, suggesting that services are often aligned with their specific needs. Group travelers, on the other hand, showed a moderate concentration of ratings in the "Good" and "Very Good" categories, representing an overall satisfying experience, albeit with clear opportunities for improvement. These findings underscore the need for tailored service strategies to meet the unique expectations of different traveler demographics.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

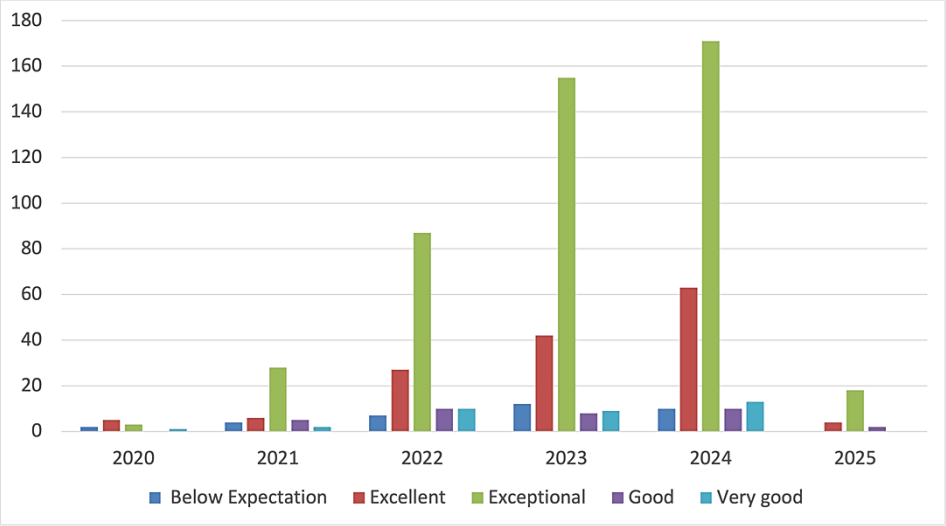

Figure 5. Expectations by Year of Visit (714 Reviews)

Figure 5 highlights the distribution of visitor expectations by year of visit, which comes from 714 reviews, showing different trends in satisfaction levels over time. The data revealed significant improvements in "Good" and "Very Good" ratings in 2023 and 2024, showing a significant increase in guest satisfaction. In contrast, previous years, such as 2020 and 2021, showed fewer reviews with more even rankings across all categories, indicating an inconsistent experience during the period. The substantial increase in "Excellent" and "Excellent" rankings in 2022 and 2024 underscores a positive shift in the quality of service or visitor perception. This trend will likely reflect the implementation of improved service strategies or recovery efforts in response to previous challenges. These findings emphasize the importance of monitoring temporal trends in satisfaction to identify periods of success and areas that need further refinement.

The distribution of visitor satisfaction levels over the years shows significant variation, reflecting trends in service quality and guest experience. In 2020 and 2021, the number of reviews was relatively low, with the percentage evenly distributed across all categories and a small percentage categorized as "Below Expectations," indicating inconsistent satisfaction during these years. A notable increase occurred in 2022, as "Good" and "Very Good" ratings accounted for 40–50% of reviews, complemented by a substantial increase in "Excellent" ratings (~20–30%), indicating an improvement in service quality and visitor satisfaction. In 2023, satisfaction rates increased, with "Good" and "Very Good" ratings dominating 60–70%, reflecting a consistent effort to meet and exceed guest expectations. The year 2024 marked the peak of satisfaction, with around 70% of reviews rated as "Good" and a significant proportion categorized as "Excellent" and "Excellent" (20–30%), highlighting a period of exceptional service delivery. Although data for 2025 is limited, the preliminary review reveals a balanced distribution with little emphasis on "Good" and "Very Good," indicating a continuation of the favorable trend. These patterns underscore the importance of continuous service improvement efforts to maintain high levels of guest satisfaction.

3. RESULTS AND DISCUSSION

Analyzing sentiment distributions and scores provides essential insights into the guest experience, highlighting satisfaction patterns and areas of improvement and aligning service quality with expectations. Positive sentiment dominates, reflecting intense service satisfaction, while neutral sentiment suggests opportunities for improvement, and negative sentiment emphasizes critical areas that need attention. These evaluations support data-driven strategies to improve service delivery and guest experience effectively.

3.1 Distribution Analysis and Score

The distribution and analysis of guest review scores provide essential insights into visitor satisfaction patterns and service quality. Analyzing the spread of review scores highlights the prevalence of specific satisfaction levels across different demographics and visitor time frames. This assessment is crucial in identifying trends, such as the dominance of favorable ratings such as "Good" and "Very Good" or the occurrence of lower scores categorized as "Below Expectations." The pattern shows the strengths and weaknesses of service delivery, which reflects the alignment of hospitality offerings with customer expectations. A detailed examination of these scores supports identifying areas that require strategic improvement while reinforcing successful practices. By systematically understanding distribution trends, these analytics facilitate data-driven decision-making to improve the guest experience and maintain competitive service standards in the hospitality industry.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

Figure 6. Overall Distribution

Figure 6 illustrates the overall distribution of sentiment derived from guest reviews, which are categorized into positive, neutral, and negative sentiments. Positive sentiment dominated the data set, reflecting a significant proportion of guest experiences aligned with expectations, demonstrating consistent satisfaction with service quality. Neutral sentiment accounts for a moderate percentage, showing instances where guest feedback reflects an average or mixed perception of the service provided. While representing the most minor proportion, negative sentiment highlights critical areas that need attention to address dissatisfaction effectively. This distribution underscores the importance of identifying patterns in guest feedback to optimize service delivery and improve customer satisfaction. Analyzing this sentiment provides actionable insights to prioritize improvements, ensuring a balanced approach to maintaining high service standards while addressing areas of concern.

The sentiment distribution in the dataset reveals significant insights into the guest experience and service quality. Positive sentiment, accounting for 60–70% of reviews, dominated the data set, demonstrating a strong alignment between guest expectations and the service provided, reflecting consistent satisfaction across most visitors. Neutral sentiment, which ranges from around 20-30%, indicates that most guests experience an average or mixed perception, highlighting areas where service delivery meets expectations but fails to provide a great experience. Negative sentiment, which represents about 10-15%, identifies specific areas of dissatisfaction, which, while limited in scale, carry critical implications for overall guest perception. This distribution underscores the need for targeted strategies to address the weaknesses highlighted in negative sentiment while leveraging the strengths reflected in positive feedback to improve overall service excellence.

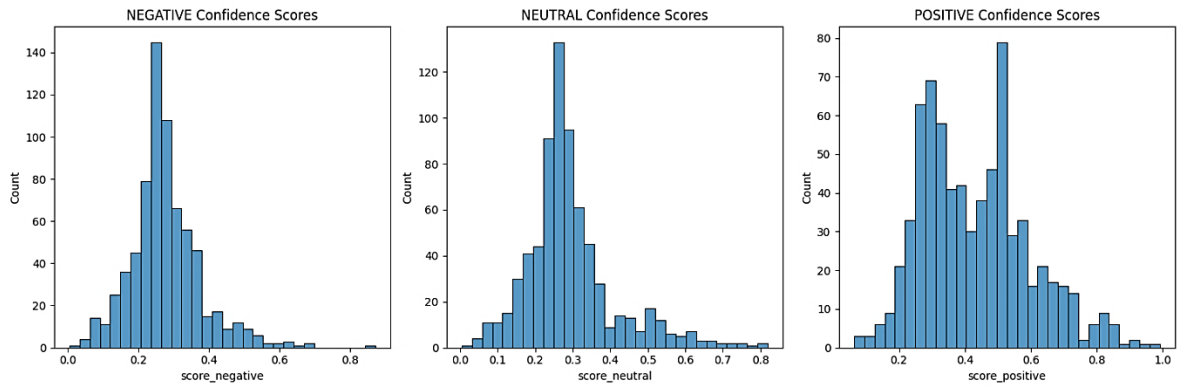

Figure 7. Score Distribution per Class

Figure 7 illustrates the distribution of confidence scores across three sentiment classes: negative, neutral, and positive, providing a detailed perspective on the model's predictive performance. Negative sentiment scores indicate a skewed distribution to the left, with most values below 0.5, indicating lower confidence in classifying reviews as very negative. In contrast, the neutral sentiment score shows a normal distribution centered around 0.5, which reflects the model's balanced approach to identifying reviews with an average or ambiguous tone. Positive sentiment scores show a right-skewed distribution, with higher frequencies observed in the range of 0.5 to 0.8, indicating a more substantial model confidence in classifying positive reviews. This distribution analysis underscores the model's tendency to give higher confidence to positive classifications while maintaining moderate accuracy in neutral predictions and a cautious approach to negative reviews. These patterns provide valuable insights to refine model performance and improve classification reliability across all sentiment categories.

The distribution of confidence scores across sentiment classes reveals different patterns that highlight the model's performance, which varies in sentiment classification. Negative sentiment indicates a skewed distribution to the left, with most scores below 0.5, reflecting the model's conservative approach to identifying very negative reviews. This cautious strategy has the potential to minimize false negatives but risks downplaying dissatisfaction. Neutral sentiment displays a near-normal distribution centered around 0.5, indicating balanced

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

confidence in classifying ambiguous or average reviews. However, this pattern allows for increased precision to reduce overlap with adjacent classes. The positive sentiment, which is characterized by a right-skewed distribution, shows an intense concentration of confidence scores in the range of 0.5 to 0.8, indicating the model's reliability in recognizing favorable reviews. Fewer scores exceeded 0.8, indicating limitations in identifying positive sentiments with maximum certainty. Overall, this analysis underscores the model's strength in classifying positive sentiment while highlighting the need for refinement in handling both neutral and negative sentiment to achieve a more balanced and robust classification framework.

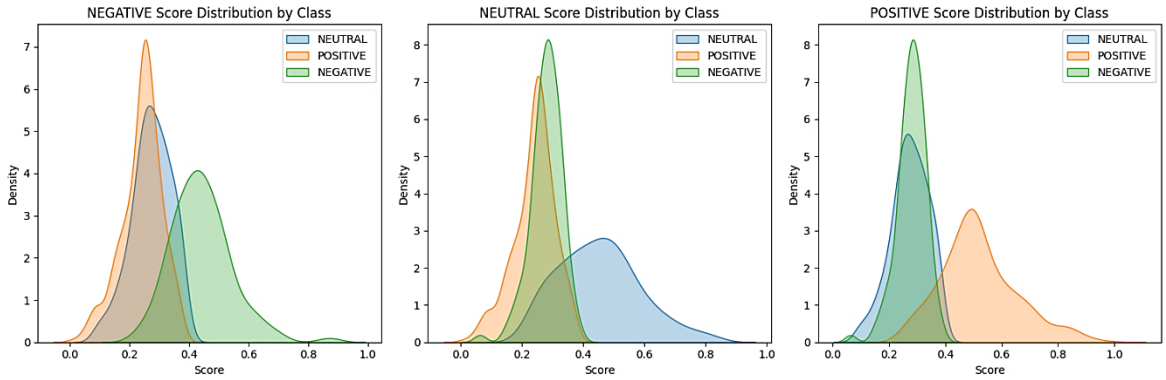

Figure 8. Comparison of Distribution Between Classes

Figure 8 provides a comparative visualization of the distribution of scores across sentiment classes, highlighting overlaps and variations in the model's classification performance. Negative classes show a left-skewed distribution, with the highest density near the lower score range, reflecting the model's careful determination of confidence for very negative reviews. The neutral class shows a near-symmetrical distribution centered around a score of 0.5, indicating the model's balanced approach and potential overlap with adjacent classes. The positive class shows a right-sloping distribution, with significant densities in the range of 0.5 to 0.8, illustrating a more substantial model confidence in identifying favorable sentiments. However, the overlapping regions between the classes, especially between the neutral and the other two classes, indicate areas where the precision of the model can be improved. This comparative analysis underscores the model's strength in identifying positive sentiment while emphasizing the need for further refinement in distinguishing between neutral and negative reviews, thereby improving the overall classification accuracy.

Analyzing the distribution of scores across sentiment classes highlights different patterns and overlaps, providing valuable insights into model performance. Negative sentiment shows a skewed distribution to the left, with confidence scores mostly below 0.5, indicating the model's cautious approach to classifying very negative reviews. The significant overlap with neutral sentiment in the range of 0.4 to 0.6 indicates a challenge in distinguishing rather negative feedback from neutral responses. Low densities above 0.6 underscore the limitations of confidently identifying highly negative reviews, potentially leading to underclassification of dissatisfaction. Neutral sentiment, which is characterized by a near-normal distribution centered around 0.5, reflects balanced confidence in classifying average or ambiguous reviews. However, overlap with negative sentiment in the lower range (0.3 to 0.5) and positive sentiment in the higher range (0.5 to 0.7) reveals difficulty separating neutral feedback from extremes, increasing the risk of misclassification. Positive sentiment indicates a right-skewed distribution, with the highest density between 0.5 and 0.8, indicating more substantial confidence in recognizing favorable reviews.

Nevertheless, a minimum density above 0.8 highlights the constraints in identifying very positive feedback with maximum certainty. Comparatively, positive sentiment shows the highest confidence and least overlap, indicating a reliable classification, while neutral sentiment faces the most significant overlap and ambiguity. Negative sentiment, with the lowest confidence overall, requires further refinement to improve the model's ability to capture dissatisfaction effectively. The findings emphasize the importance of addressing overlapping regions and increasing confidence levels for more accurate sentiment classification.

3.2 Model Performance Analysis

The model's performance evaluation highlights its ability to classify sentiment into three categories: positive, neutral, and negative, with varying degrees of precision and confidence. The positive sentiment classification shows the strongest performance, characterized by a high confidence score and minimal classification errors, which indicates the model's reliability in recognizing favorable reviews. The neutral sentiment classification, while balanced, reveals challenges due to significant overlap with positive and negative classes, especially on mid-range confidence scores, indicating the need for increased precision in distinguishing ambiguous sentiments. On the other hand, the negative sentiment classification indicates the lowest level of confidence, with most scores grouping in a lower range, reflecting the model's cautious approach to identifying dissatisfaction and

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

pointing to the risk of underclassification. These patterns suggest that while the model excels at identifying positive sentiment, targeted refinements are needed to improve its ability to classify neutral and negative feedback accurately. Strengthening these aspects will contribute to a more balanced and robust sentiment analysis framework, ensuring excellent reliability across all sentiment categories.

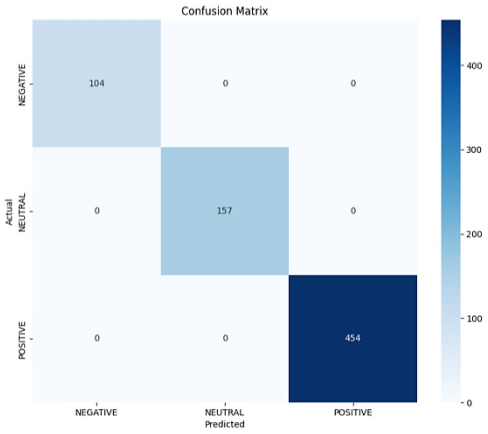

Figure 9. Confusion Matrix

Figure 9 presents a confusion matrix, which illustrates the performance of sentiment classification models in three categories: negative, neutral, and positive. The matrix highlights many correctly classified positive reviews, as reflected by the dominant values in the bottom right cell, demonstrating the model's strong performance in identifying favorable sentiments. Neutral sentiment indicates a moderate level of correct classification, indicated by the intermediate range values in the middle cell, while also revealing some misclassification into positive or negative categories, indicating overlap and ambiguity in neutral predictions. Negative sentiment, represented by the upper left cell, shows a less correct classification than the positive and neutral categories, highlighting the challenge in identifying dissatisfaction. The sparse distribution of values outside the diagonal cells suggests that this model is generally effective but faces difficulty eliminating classification errors. This analysis underscores the model's reliability in identifying positive sentiment while demonstrating the need for refinement in distinguishing between neutral and negative sentiment to improve accuracy and robustness.

Confusion matrix analysis reveals the model's performance in classifying positive, neutral, and negative sentiments with varying degrees of accuracy. Positive sentiment achieved the highest accuracy, with 424 reviews classified correctly and only eight incorrectly classified as neutral. This demonstrates the model's reliability in identifying favorable feedback and its strong alignment with positive sentiment recognition. On the other hand, neutral sentiment showed moderate performance, with 127 reviews correctly classified but nine incorrectly classified as positive and eight as unfavorable, reflecting the inherent challenge of distinguishing neutral reviews due to overlapping with adjacent classes. Negative sentiment showed the lowest accuracy, with 104 reviews correctly classified and nine incorrectly classified as neutral, underscoring the model's conservative approach and struggling to identify strong negative sentiment with greater precision. These patterns suggest that although the model performs very well with positive sentiment, refinements are needed to improve its precision and sensitivity in classifying neutral and negative sentiment. Targeted handling overlap and increased ambiguity will significantly improve model robustness and performance.

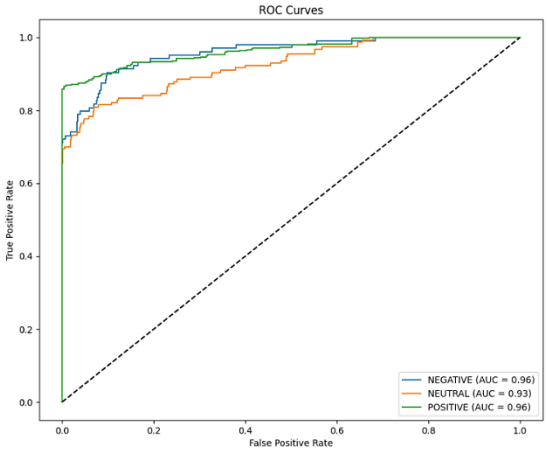

Figure 10. ROC Curve

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

Figure 10 illustrates the Receiver Operating Characteristics (ROC) curve for model classification performance across negative, neutral, and positive sentiment classes. The curve represents the trade-off between the accurate positive level (sensitivity) and the false positive level for each sentiment class. The Under the Curve Area (AUC) value is 0.96 for negative, 0.93 for neutral, and 0.96 for positive, indicating strong overall performance. The negative and positive sentiment classes show near-perfect AUC scores, reflecting the model's ability to distinguish between positive and negative examples accurately. However, the neutral sentiment class showed a slightly lower AUC, suggesting a challenge in balancing sensitivity and specificity, likely due to overlapping characteristics with adjacent classes. The overall performance, as reflected by the AUC value, shows the robustness of the model, especially in identifying different positive and negative sentiments. Strengthening the precision in neutral classification will further increase the model's efficacy in handling diverse sentiment data.

The ROC analysis highlights the model's performance in differentiating between negative, neutral, and positive sentiments, with significant variations in effectiveness. Negative sentiment reached an AUC of 0.96, reflecting excellent classification accuracy as the curve approached the upper left corner, indicating high true and minimal false positives. Neutral sentiment, with an AUC of 0.93, showed slightly lower performance as the curve deviated from the ideal shape, revealing challenges in separating neutral reviews from adjacent classes, especially in ambiguous cases with overlapping characteristics. Positive sentiment also reached an AUC of 0.96, demonstrating strong reliability in identifying favorable feedback, with the curve maintaining an optimal trajectory and minimal false classifications. A comparison of these classes shows that while the classification of negative and positive sentiment shows near-perfect performance, the neutral class lags slightly behind, reflecting the complexity inherent in the differences. Strengthening the model's ability to distinguish neutral sentiment will improve its robustness and ensure excellent reliability in sentiment classification.

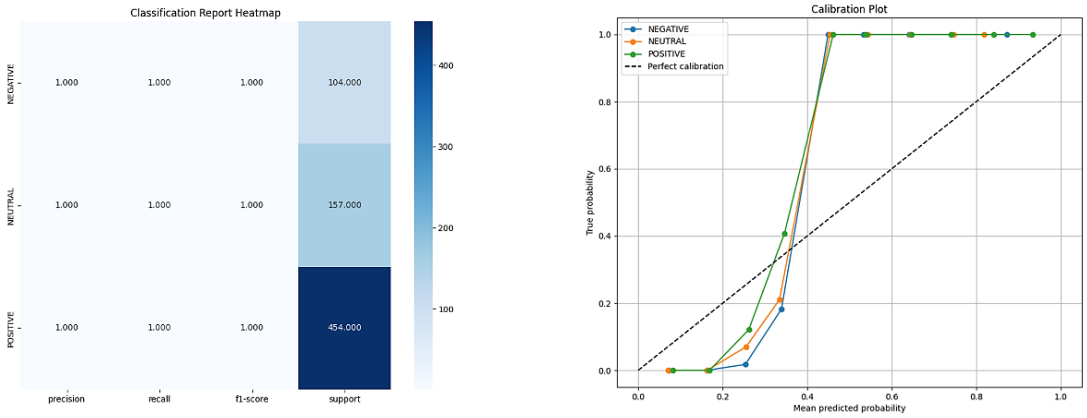

Figure 11. Classification Report and Calibration Plot

Figure 11 illustrates a heatmap of the classification report and calibration plot, providing insight into the model classification metrics and probability calibration. The classification report heatmap presents key evaluation metrics, including precision, draws, f1 scores, and support for each sentiment class. The model shows high positive and negative sentiment performance, as reflected in the high precision and drawdown values. At the same time, neutral classes show relatively lower metrics, indicating challenges in accurately capturing and classifying ambiguous reviews. The calibration plot assesses the alignment between the predicted probability and the actual outcome. Positive and negative classes indicate well-calibrated probabilities following a diagonal line, indicating that the model's confidence aligns with its predictive accuracy. However, neutral classes sometimes show slight deviations, indicating excessive or lack of confidence. This analysis highlights the model's power in handling different positive and negative sentiments while identifying the need to improve the precision and calibration of neutral sentiment predictions. Improving these aspects will contribute to a more reliable and balanced classification system.

The data comprehensively evaluates the model's classification performance, combining classification report heatmaps and calibration plots to assess precision calibration, drawdown, F1 score, support, and probability for sentiment classes. The positive sentiment class achieved outstanding metrics, with a precision of 0.98, a memory of 0.97, and an F1 score of 0.98, supported by a substantial review volume of 432, which reflects the model's reliability in identifying favorable sentiment with minimal classification errors. Neutral sentiment showed moderate performance, with precision, memory, and an F1 score of 0.87, based on small support from 144 reviews, indicating challenges in accurately capturing ambiguous sentiment. Negative sentiment performed strongly, with precision, recall, and an F1 score of 0.92, although the most negligible support of 113 reviews highlighted fewer instances of dissatisfaction. The calibration plot reveals that the positive and negative classes align with the predicted probabilities following the diagonal line. In contrast, the neutral class showed slight

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

deviations in the mid-range probabilities, indicating a lack of confidence in some predictions. These findings emphasize the model's robust ability to handle both positive and negative sentiment while highlighting the need for targeted refinements in the neutral sentiment classification to improve overall balance and reliability.

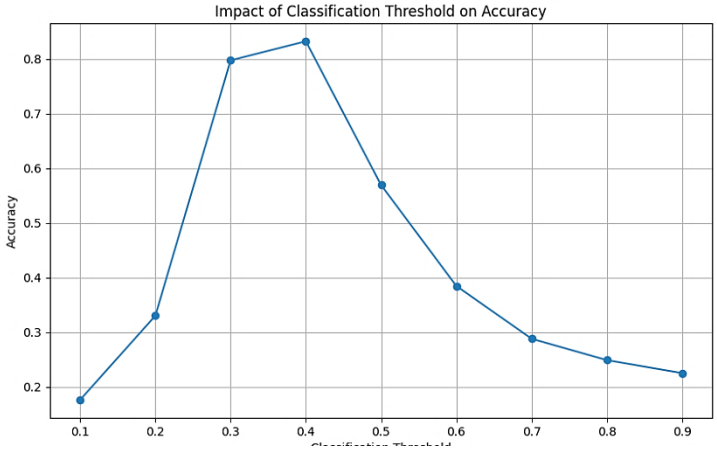

Figure 12. Impact of Classification Thresholds on Accuracy

Figure 12 illustrates the relationship between classification thresholds and model accuracy, showing how threshold variations affect model performance. The chart reveals that the accuracy peaked at the threshold of 0.4, reaching a high value of around 0.85, indicating the optimal balance between precision and retraction. The accuracy decreased significantly as the threshold moved away from this optimal value, especially outside of **0.5**, where it continued to drop below 0.3, highlighting a decline in the model's ability to classify sentiment at higher thresholds effectively. Similarly, a threshold below 0.3 results in lower accuracy, as the model is likely to become less precise due to the increase in false positives. This trend shows that precisely selected thresholds are critical to maximizing model accuracy. This analysis underscores the importance of refining classification thresholds to achieve optimal performance, ensuring a balance between correctly identifying sentiment and minimizing misclassification.

3.2 Discussion

Examining the sentiment distribution in guest reviews reveals a predominance of positive feedback, which shows a substantial alignment between service delivery and customer expectations. This underscores a consistent level of satisfaction across different aspects of hospitality. Examples of neutral sentiment show that while basic expectations are met, distinctive or memorable service features are absent, indicating the potential for strategic innovation to improve the customer experience. While negative sentiment represents a smaller portion of the data set, it provides essential insights into specific service deficiencies that require immediate attention to reduce dissatisfaction and prevent reputational risk. This distribution analysis highlights the need to strengthen existing strengths while addressing operational deficiencies to maintain a competitive advantage. A comprehensive understanding of these patterns facilitates the design of targeted interventions, ensuring that service offerings meet and exceed customer expectations, thereby fostering long-term satisfaction and loyalty in an increasingly competitive industry landscape.

The model's performance assessment highlights the critical forces in classifying positive sentiment, which is characterized by a high confidence level and a minimal classification error rate, demonstrating the model's reliability in identifying a favorable guest experience. Nonetheless, the classification of neutral sentiment presents significant challenges, mainly due to overlapping boundaries with adjacent categories, which reflects the ambiguity inherent in specific feedback. Similarly, the classification of negative sentiment shows limited accuracy, which stems from cautious prediction thresholds and reduced confidence in identifying dissatisfaction being voiced. These issues point to improved methodological refinements, especially in addressing ambiguity in neutral classifications and increasing sensitivity to detecting nuanced negative sentiments. The overall precision and interpretability of the model can be significantly improved by optimizing these aspects, allowing for a more balanced and robust sentiment analysis framework across different categories of guest feedback.

The distribution of confidence scores reveals significant variations across sentiment classes, with positive sentiment indicating high reliability and consistency, demonstrating the model's strength in identifying favorable feedback. In contrast, neutral sentiment scores show a balanced but overlapping distribution, especially with negative scores, reflecting the ambiguity inherent in classifying nuanced responses. This overlap suggests that the model faces the challenge of clearly distinguishing between moderately critical and neutral feedback, potentially reducing the accuracy of its predictions. Addressing these issues through advanced contextual embedding or integrating hybrid modeling techniques is essential to improve the model's ability to interpret and classify sentiment accurately. Such refinements will ensure improved performance in handling ambiguous

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

sentiment and foster a more reliable and adaptable sentiment analysis framework capable of effectively capturing the complexities of guest perceptions.

Insights gained from the distribution of sentiment and classification performance provide substantial value for strategic decision-making in hospitality management, especially in improving service quality and guest satisfaction. Positive feedback strengthens the service delivery framework, emphasizing areas that contribute to high customer satisfaction and guarantee reinforcement. In contrast, neutral and negative sentiment highlights essential aspects of the service experience that require targeted improvements to address unmet expectations and reduce dissatisfaction. These insights offer a strong foundation for implementing data-driven strategies to optimize operational efficiency and align services with evolving customer demands. By leveraging this information, hospitality providers can devise tailored interventions to improve the guest experience and strengthen customer loyalty, ensuring continued competitiveness in an ever-changing market landscape.

4. CONCLUSION

Applying the Multilingual BERT model has significantly improved the accuracy of sentiment classification in hotel guest reviews, especially in recognizing positive sentiment with 98% precision, 97% recall, and 98% F1 score, as observed in 432 correctly classified reviews. This study uses the Knowledge Discovery in Databases (KDD) methodology, which includes data selection, preprocessing, transformation, sentiment classification, and performance evaluation. The dataset consists of 715 hotel reviews from Qubika Boutique Hotel, obtained from the Agoda platform, ensuring a wide range of guest sentiment. Despite the increase in recognition of positive sentiment, challenges remain in classifying neutral sentiment, which achieves a lower precision of 87%, with 127 cases correctly classified, and negative sentiment, where accuracy remains at 92%, with 104 reviews correctly identified. The overlap in confidence scores, especially in the range of 0.4-0.6 between neutral and negative sentiment, underscores the need for enhanced contextual embedding and hybrid modeling techniques to improve the reliability of the classification. Sentiment distribution analysis revealed that 60-70% of reviews reflect a positive experience, 20-30% are neutral, and 10-15% indicate dissatisfaction, necessitating targeted service improvements. These findings provide actionable insights for hospitality management, enabling data-driven strategies to improve customer satisfaction and operational efficiency. Future research should focus on refining the interpretability and adaptability of models by incorporating advanced deep-learning architectures, expanding multilingual datasets, and integrating real-time sentiment analysis. Strengthening these aspects will contribute to a more robust and measurable sentiment analysis framework, ensuring greater precision in capturing guest experiences and optimizing service strategies in the hospitality industry.

REFERENCES

[1] O.A. George and CMQ Ramos, "Sentiment analysis applied to tourism: exploring tourist-generated content in the case of health tourism destinations," *J. Spa International Health*, vol. 7, no. 2, hlm. 139–161, 2024, doi: 10.1080/24721735.2024.2352979.

[2] A. Ameur, S. Hamdi, and S. Ben Yahia, "An enhanced multilabel learning approach for the detection of Arabic aspect categories of hotel reviews," *Computing. Intell.*, Vol. 40, No. 1, 2024, doi:10.1111/coin.12609.

[3] Y. Wu, J. Wang, Y. Xia, Q. Li, and Y. Pan, "Sensing the distribution of hotel customers and the variation in their sentiment using online travel agency data: the case of a Shanghai star hotel," *Ann. GIS*, vol. 30, no. 3, hlm. 323–343, 2024, doi: 10.1080/19475683.2024.2335976.

[4] J. Wang, "Hotel Room Experience Design Based on Virtual Reality Technology," *A. Electricity. System.*, vol. 20, no. 1, hlm. 206–218, 2024, doi: 10.52783/jes.677.

[5] Y. A. Singgalen, S. Y. Wahyuningtyas, Y. E. Widodo, M. N. A. Dasra, and R. W. Setiawan, "Discovery of Knowledge in Databases for Improving Hotel Service Quality Through a Data Mining Approach," *J. Teor. Inf. Technol App.*, vol. 102, no. 24, pp. 9004–9020, 2024, [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85213991405&origin=inward

[6] J. L. Nicolau, Z. Xiang, and D. Wang, "Daily online review sentiment and hotel performance," *Int. J. Contemporary. Manag Hospital.*, vol. 36, no. 3, hlm. 790–811, 2024, doi: 10.1108/IJCHM-05-2022-0594.

[7] M.G. Gîngioveanu Lupulescu, V.M. Dincă, SD Taranu, and B.A. Blănuță, "Data-Driven Insights from 10,000 Reviews: Driving Sustainability through Rapid Adaptation to Guest Feedback," *Defend.* , vol. 16, no. 7, 2024, doi: 10.3390/SU16072759.

[8] JT Hsueh and S.H. Hsu, "Turning negative reviews into operational insights: The role of ABSS-GPT in informing hotel decisions," *J. Decis. System.*, 2024, doi: 10.1080/12460125.2024.2428977.

[9] M.J. Sánchez-Franco and S. Rey-Tienda, "The role of user-generated content in tourism decision-making: a case study of Andalusia, Spain," *Manag. Results.*, Vol. 62, No. 7, hlm. 2292–2328, January 2024, Yogurt: 10.1108/MD-06-2023-0966.

[10] S. Gupta and R. Jaiswal, "How We Can Improve Hospitality Excellence for Sustainable Development Using Machine Learning," *J. Hosp. Tour. Emerged.*, 2024, doi: 10.1080/10963758.2024.2420267.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

[11] LC Cheng, H.Y. Huang, and Y.W. Huang, "A multi-task China aspect-based sentiment analysis framework for service improvement: a case study on BNB reviews," *Electron. Comes. Res.*, 2024, doi: 10.1007/s10660-024-09871-0.
[12] N.K. Boparai, H. Aggarwal, and R. Rani, "Analyzing review fuzzy semantics for multi-criteria recommendations," *Data Knowledge. Eng.*, Vol. 152, 2024, doi: 10.1016/J. Donork.2024.102314.
[13] F. Jeribi, U. Perumal, and M. H. Alhameed, "A Recommendation System for Sustainable Day and Night Cultural Tourism Using Recurrent Neural Networks Centered on Average Marked Errors for Riyadh Historic Sites," *Defend.* , vol. 16, no. 13, 2024, doi: 10.3390/SU16135566.
[14] S. Bhowmik, R. Sadik, W. Akanda, and J. R. Pavel, "Sentiment analysis with hotel customer reviews using FNet," *Cow. Electricity. Eng. Informatics*, vol. 13, no. 2, no. 1298–1306, 2024, doi: 10.11591/eei.v13i2.6301.
[15] Y. Andriyana *et al.*, "Spatial Durbin Model with Expansion Using the Casetti Approach: A Case Study of Rainfall Prediction in Java, Indonesia," *Mathematics*, vol. 12, no. 15, 2024, doi: 10.3390/math12152304.
[16] G.D. Mendonça, S.R. de M. Oliveira, OFlima, and PTV de Resende, "Intelligent algorithms applied to air transport delay prediction," *Int. J. Phys. Distrib. Logistik. Manag.*Vol. 54, No. 1 hlm. 61–91, January 2024, doi:10.1108/IJPDLM-10-2022-0328.
[17] S. Khlamov, V. Savanevych, T. Trunova, Z. Deineko, O. Vovk, and R. Gerasimenko, "Automatic Data Mining of Reference Stars from Astronomical CCD Frames," *CEUR Workshop Proceedings*, vol. 3668. pp. 83–97, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85191652348&origin=inward
[18] K. Talebi, Z. Torabi, and N. Daneshpour, "An ensemble model based on CNN and LSTM for dropout prediction in MOOC," *Application Expert System.*, vol. 235, 2024, doi: 10.1016/j.eswa.2023.121187.
[19] D. Srivastava, R. Singh, C. Chakraborty, S. K. Maakar, A. Makkar, and D. Sinwar, "A framework for detecting cyberattacks with intrusion detection dataset classification," *Microprocesses. Microsystems.*, vol. 105, 2024, doi: 10.1016/j.micpro.2023.104964.
[20] M. N. Razali, S. A. Manaf, R. B. Hanapi, M. R. Salji, L. W. Chiat, and K. Nisar, "Improving the Classification of Minority Sentiment in Gastronomic Tourism: A Hybrid Sentiment Analysis Framework with Data Augmentation, Feature Engineering, and Business Intelligence," *IEEE Access*, vol. 12, no. December 2023, hlm. 49387–49407, 2024, doi: 10.1109/ACCESS.2024.3362730.
[21] T. Mahmud, M. Ptaszynski, and F. Masui, "A Complete Study of Machine Learning and Deep Learning Methods for the Detection of Multilingual Cyberbullying in Bangla and Chittagonian Texts," *Electron.*, Vol. 13, No. 9, 2024, doi: 10.3390/es13091677.
[22] D. Gupta *et al.*, "True and Fraudulent Hotel Reviews Based on Deep Learning," *Defend.* , vol. 16, no. 11, 2024, doi: 10.3390/SU16114514.
[23] M. Ijaz, N. Anwar, M. Safran, S. Alfarhood, T. Sadad, and Imran, "Domain adaptive learning for multi-domain sentiment classification on big data," *PLoS One*, Vol. 19, no. 4 April 2024, Yogurt: 10.1371/Journal.Pone.0297028.
[24] N. Habbat, H. Anoun, L. Hassouni, and H. Nouri, "Hotel Demand Forecasting through Booking Comments Using Sentiment Analysis and Topic Modeling Techniques," *Advancements in Science, Technology, and Innovation*. pp. 113–122, 2024. doi: 10.1007/978-3-031-46849-0_13.
[25] N. Habbat and H. Nouri, "Unlocking travel narratives: a blend of deep learning composing ensemble and neural topic modeling for enhanced analysis of tourism commentary," *Soc. Netw. Anal. Min.*, Vol. 14, No. 1, 2024, doi:10.1007/S13278-024-01256-3.
[26] W. Jin *et al.*, "Improving rural B&B management through machine learning and evolutionary games: A case study of rural revitalization in Yunnan, China," *PLoS One*, vol. 19, no. 3 March, 2024, doi: 10.1371/journal.pone.0294267.
[27] M. Maryamah, G. Wilsen, C. T. Suhalim, R. Septiana, A. Fajar, and M. I. Solihin, "Hybrid Information Retrieval with Masked and Permuted Language Modeling (MPNet) and BM25L for Indonesian Drug Data Collection," in *KST 2024 - 16th International Conference on Smart Knowledge and Technology2024, pp. 242–247. doi:10.1109/KST61284.2024.10499674.*
[28] A. Riyadi, M. Kovacs, U. Serdült, and V. Kryssanov, "IndoGovBERT: A Domain-Specific Language Model for Processing Indonesian Government SDG Documents," *Big Data Cogn. Computing.*, Vol. 8, No. 11, 2024, doi: 10.3390/BDCC8110153.
[29] H. Al-Jarrah, M. Al-Smadi, M. Hammad, and F. Shannaq, "Using Deep Learning Techniques to Detect Hate and Abusive Language in Arabic Tweets," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, hlm. 553–569, 2024, doi:10.22266/ijies2024.1031.43.
[30] E. Raja, B. Soni, and S. K. Borgohain, "Harnessing heterogeneity: A multi-embedding ensemble approach to detecting fake news in Dravidian language," *Computing. Electricity. Eng.*, vol. 120, 2024, doi: 10.1016/j.compeleceng.2024.109661.
[31] M. E. Hassan, M. Hussain, I. Maab, U. Habib, M. A. Khan, and A. Masood, "Detection of Sarcasm in Urdu Tweets Using Deep Learning and a Transformer-Based Hybrid Approach," *IEEE Access*, vol. 12, hlm. 61542–61555, 2024, doi: 10.1109/ACCESS.2024.3393856.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 2, February 2025, Page 508-520
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i2.6870

[32] K. Kim *et al.*, "A Multifaceted Natural Language Processing Task-Based Evaluation of Two-Way Encoder Representations from Transformers Models for Bilingual Clinical Records (Korean and English): Algorithm Development and Validation," *JMIR Med. Informatics*, vol. 12, 2024, doi: 10.2196/52897.

[33] A. Dhakshina Moorthy, D. Kavitha, R. Logeshwaran, N.V. Vishnu Kumar, and V. Karthick, "PSFAS: A Progressive Student Feedback Analysis System to Improve Teaching Learning with Intelligent Processing of Open Responses," *J. Appl. Res. Tinggi. Emerged.*2024, Doi: 10.1108/Jarhe-04-2024-0157.

[34] D. Karpov and M. Burtsev, "Monolingual and Cross-Language Knowledge Transfer for Topic Classification," *J. Math. Sci. (United States)*, vol. 285, no. 1, hlm. 36–48, 2023, doi:10.1007/s10958-024-07421-5.

[35] KA Alshaikh, OA Almatrafi, and YB Abushark, "A BERT-Based Model for Aspect Based Sentiment Analysis to Analyze Arabic Open Survey Responses: A Case Study," *IEEE Access*, vol. 12, no. January, hlm. 2288–2302, 2024, doi: 10.1109/ACCESS.2023.3348342.

[36] A.H. Aljammal, I. Al-Oqily, M. Obiedat, A. Qawasmeh, S. Taamneh, and F.I. Wedyan, "Detection of anomalous intrusion using machine learning-IG-R based on NSL-KDD datasets," *Cow. Electricity. Eng. Informatics*, vol. 13, no. 6, no. 4466–4474, 2024, doi: 10.11591/eei.v13i6.7308.

[37] B. Al-Fuhaidi, Z. Farae, F. Al-Fahaidy, G. Nagi, A. Ghallab, and A. Alameri, "Anamal-Based Intrusion Detection System in Wireless Sensor Networks Using Machine Learning Algorithms," *Application. Computing. Intell. Soft computing.*, vol. 2024, 2024, doi: 10.1155/2024/2625922.