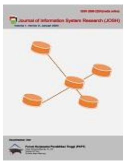


Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 976−986
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6505

Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data

Yerik Afrianto Singgalen*

Faculty of Business Administration and Communication, Tourism Study Program, Atma Jaya Catholic University of Indonesia, Jakarta
Jl. Jend. Sudirman No.51 5, RT.004/RW.4, Karet Semanggi, Setiabudi District, South Jakarta City, Special Capital Region of Jakarta, Indonesia
Email: yerik.afrianto@atmajaya.ac.id
Correspondence Author Email: yerik.afrianto@atmajaya.ac.id

Submitted: **23/12/2024;** Accepted: **03/01/2025;** Published: **05/01/2025**

Abstract−This study investigates the performance of a sentiment classification model leveraging IndoBERT to analyze Indonesian hotel review data. Sentiment analysis is crucial for extracting actionable insights from customer reviews, yet challenges such as linguistic diversity and imbalanced datasets complicate accurate classification. The dataset comprises 90% Positive, 5% Neutral, and 5% Negative sentiments, reflecting significant class imbalance. A fine-tuned IndoBERT model was trained over three epochs, with performance assessed using metrics such as accuracy, precision, recall, F1-score, confusion matrices, and ROC and Precision-Recall curves. The results indicate high global accuracy (92.52%) and robust performance for the Positive class (F1-score: 96.09%, AUC: 0.90). However, significant limitations were observed for minority classes, with the Neutral class achieving precision, recall, and F1-scores of 0.00, and the Negative class obtaining a low F1-score of 28.57%. These findings underscore the influence of dataset imbalance, where the dominance of the Positive class skews model predictions. Future research should explore techniques such as oversampling SMOTE, reweighting loss functions, or hybrid architectures to mitigate imbalance and improve performance across all sentiment categories. This research contributes to advancing sentiment classification methodologies for Indonesian text, offering practical implications for enhancing customer feedback analysis in the hospitality industry.

Keywords: IndoBERT; Sentiment Analysis; Hotel Reviews; Class Imbalance; Model Performance

1. INTRODUCTION

The rapid growth of online reviews has elevated the importance of sentiment classification in various domains, including the hospitality industry, where customer feedback significantly influences business performance. Sentiment analysis, particularly in Indonesian hotel review data, presents unique linguistic challenges due to the language's complexity and informal nature [1]. IndoBERT, a pre-trained language model optimized for Bahasa Indonesia, holds promise for addressing these challenges by leveraging contextual embeddings tailored to the language's intricacies [2]. Evaluating the model's performance in sentiment classification advances understanding of its capabilities and reveals its limitations when applied to domain-specific datasets [3]. Employing IndoBERT in this context raises questions about its adaptability to diverse review structures, varying levels of formality, and the inclusion of slang or regional dialects [4]. Through systematic analysis, it becomes evident that the model's strength lies in capturing semantic nuances, although certain tokenization and data preprocessing methods are crucial to optimizing its accuracy. Therefore, this investigation underscores the necessity of refining techniques for sentiment classification to enhance accuracy and applicability in real-world settings.

The increasing reliance on sentiment analysis in decision-making processes across various industries highlights the critical need for accurate and efficient methodologies, particularly in linguistically diverse contexts such as Indonesian hotel reviews. The complexity of the Indonesian language, with its rich morphology, colloquial expressions, and regional dialects, poses significant challenges for traditional sentiment classification approaches [5]. Addressing these challenges is essential to ensure that sentiment analysis tools provide meaningful insights, especially in a sector where customer perceptions directly impact reputation and revenue [6]. Leveraging advanced models like IndoBERT offers a transformative opportunity to improve classification accuracy by capturing nuanced linguistic patterns [7]. Such advancements hold the potential to bridge gaps in existing methodologies, enhancing both the scalability and applicability of sentiment analysis in Indonesia's dynamic market. Therefore, this research is imperative to develop more robust analytical frameworks that can accommodate linguistic and contextual complexities and foster innovation and precision in the field.

This research aims to evaluate and enhance the performance of IndoBERT for sentiment classification in Indonesian hotel review datasets, addressing the linguistic complexities unique to this domain. By focusing on the model's ability to analyze diverse textual patterns, the study seeks to bridge gaps in existing sentiment analysis frameworks that often struggle with informal language, regional dialects, and colloquial expressions. The exploration of IndoBERT's contextual embeddings is expected to reveal its potential to accurately interpret sentiment nuances, thus contributing to more reliable sentiment analysis outcomes. A systematic examination of preprocessing techniques, tokenization strategies, and fine-tuning processes forms the foundation for optimizing the model’s performance. Ultimately, this research aims to establish a robust framework that improves sentiment classification in Indonesian and sets a benchmark for future advancements in natural language processing applications within low-resource languages.

Previous studies on sentiment analysis have extensively explored the application of machine learning and deep learning models, demonstrating substantial advancements in text classification across various languages. However, existing research often focuses on widely spoken global languages, leaving a notable gap in addressing low-resource languages such as Indonesian, particularly in domain-specific datasets like hotel reviews [8], [9]. While some efforts have been made to apply pre-trained language models to Indonesian text, many of these studies overlook the challenges posed by informal language, regional dialects, and idiomatic expressions prevalent in user-generated content [10], [11]. This limitation underscores the necessity of examining models like IndoBERT, explicitly designed for Bahasa Indonesia, in domain-specific sentiment classification tasks. Addressing this gap enriches the literature by extending the applicability of advanced natural language processing techniques and contributes to developing more inclusive and contextually accurate sentiment analysis frameworks.

This research offers significant theoretical contributions by advancing the understanding of how pre-trained language models, such as IndoBERT, perform in sentiment classification tasks within linguistically diverse and informal datasets. This study enhances the foundational knowledge of natural language processing in low-resource languages by uncovering the model's capacity to handle the complexities of the Indonesian language, including regional variations, idiomatic expressions, and colloquialisms. The practical implications are equally profound, as the findings provide actionable insights for deploying sentiment analysis tools in real-world applications, particularly in the hospitality industry, where customer feedback is pivotal. Through systematic evaluation and refinement of IndoBERT’s application, this research delivers a scalable framework that improves the precision and adaptability of sentiment classification systems. These outcomes inform future studies and support businesses and developers in implementing more effective sentiment analysis solutions tailored to the Indonesian market’s specific linguistic and contextual needs.

The novelty of this research lies in its focus on leveraging IndoBERT to address the unique linguistic challenges of sentiment classification within Indonesian hotel review datasets. Unlike previous studies that primarily target global languages or apply generalized models without tailoring to specific linguistic contexts, this study delves into the intricacies of Indonesian, encompassing informal expressions, regional variations, and domain-specific terminology [12]–[15]. By systematically evaluating and optimizing IndoBERT's performance for sentiment analysis, this research provides insights that extend beyond conventional applications, highlighting the adaptability of pre-trained language models in low-resource languages. This approach bridges a critical gap in natural language processing. It introduces a refined framework that enhances sentiment classification systems’ accuracy and contextual relevance tailored for linguistically diverse datasets [16]–[18]. Such advancements underscore the significance of localized adaptations in achieving scalable and precise analytical solutions.

2. RESEARCH METHODOLOGY

2.1 Related Works

Numerous studies have explored sentiment classification, leveraging both traditional machine learning techniques and modern deep learning frameworks, with notable advancements in accuracy and efficiency. Pre-trained language models such as BERT have gained prominence due to their ability to capture contextual relationships in textual data. However, most implementations are centered on high-resource languages, leaving languages like Indonesian underrepresented in empirical evaluations [19], [20]. Efforts to address this gap include adapting multilingual models and the development of language-specific architectures such as IndoBERT, which is tailored to the unique characteristics of Bahasa Indonesia [21]. Despite these advancements, limited attention has been given to domain-specific applications, particularly in analyzing informal and highly variable datasets like hotel reviews. This creates an opportunity to investigate IndoBERT’s effectiveness in addressing these challenges, thereby contributing to the growing body of literature on natural language processing for low-resource languages while demonstrating the potential for further innovation in sentiment analysis methodologies.

IndoBERT, a pre-trained language model specifically developed for Bahasa Indonesia, has emerged as a promising tool for sentiment analysis, particularly in handling the linguistic complexities of Indonesian reviews. Its ability to process contextual embeddings tailored to the unique syntax and semantics of the language makes it highly effective for tasks involving informal expressions, regional dialects, and nuanced sentiment patterns [22]. This capability positions IndoBERT as a superior alternative to generic or multilingual models, which often fail to fully capture the subtleties of Indonesian textual data [23]. Applying IndoBERT to sentiment analysis has significantly improved classification accuracy and contextual understanding, particularly in domains where user-generated content exhibits high variability. Such outcomes demonstrate the model's potential and highlight the importance of integrating language-specific adaptations in sentiment analysis frameworks to enhance precision and relevance in real-world applications.

A significant research gap exists in applying advanced natural language processing models to sentiment analysis in low-resource languages, mainly Indonesian, within domain-specific contexts. While substantial progress has been achieved with pre-trained models like BERT and its adaptations, many studies focus on high-resource languages or multilingual datasets, often overlooking the specific challenges posed by informal, colloquial, and context-dependent language in Indonesian user-generated content [24]. This limitation restricts the

This Journal is licensed under a Creative Commons Attribution 4.0 International License

generalizability and effectiveness of existing sentiment classification frameworks when applied to datasets such as hotel reviews characterized by diverse linguistic features [24]–[29]. Addressing this gap is essential to ensure that models can accurately interpret the unique complexities of Indonesian text, thereby advancing both theoretical and practical approaches in sentiment analysis. Such efforts contribute to the broader goal of creating more inclusive and adaptable natural language processing solutions across varied linguistic landscapes.

Future research is encouraged to explore further integrating advanced contextual language models with domain-specific enhancements to improve sentiment classification in low-resource languages like Indonesian. Expanding the focus to include diverse datasets incorporating various informal linguistic patterns, regional dialects, and specialized terminologies would provide a deeper understanding of the model's adaptability [30]. Investigating the interplay between pre-trained models and innovative preprocessing techniques, such as hybrid tokenization or dynamic embedding strategies, is likely to yield significant improvements in accuracy and efficiency. Moreover, comparative studies that assess the performance of IndoBERT against other emerging language-specific models could offer valuable insights into optimizing sentiment analysis frameworks. Such initiatives can drive innovation in natural language processing, ultimately bridging existing gaps and paving the way for more inclusive and robust analytical solutions tailored to complex linguistic environments.

2.2 Research Framework

The research framework outlines the sequential processes necessary to achieve the study's objectives, ensuring methodological rigor and clarity. It typically begins with the data phase, encompassing data collection, preprocessing, and exploratory analysis to establish a robust foundation for subsequent modeling efforts. The model phase then focuses on the sentiment classification model's iterative development, training, evaluation, and optimization, aiming to enhance its precision and adaptability to complex linguistic patterns. The final phase, analysis, synthesizes the outcomes through comprehensive results analysis and systematic documentation, providing actionable insights and supporting the study's contribution to the field. By integrating these phases cohesively, the framework ensures that each step is aligned with the research objectives while maintaining flexibility to address challenges and refine approaches. Such a design not only fosters a logical flow of activities but also supports the generation of reliable and impactful outcomes.

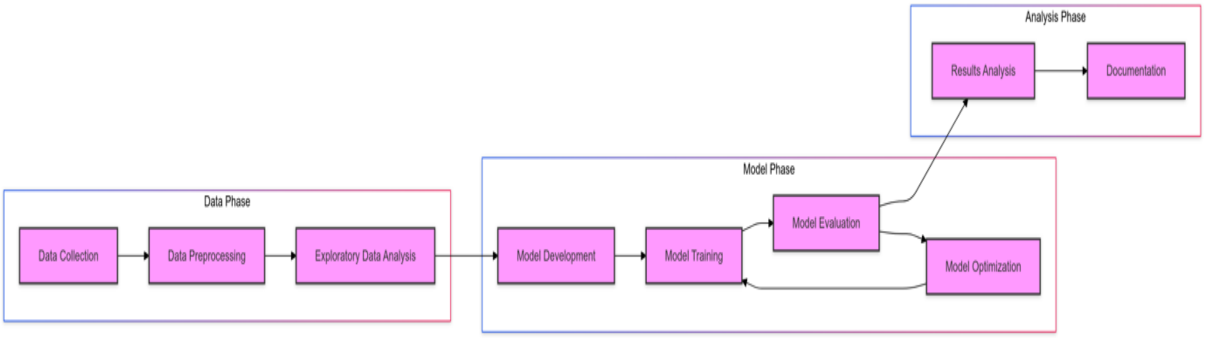


Figure 1. Research Framework

Figure 1 illustrates a structured research framework designed to systematically guide the study through its key phases: data, model, and analysis. The data phase begins with data collection, followed by preprocessing and exploratory analysis to ensure the dataset is prepared and meaningful for subsequent stages. The model phase involves iterative steps, including model development, training, evaluation, and optimization, all aimed at refining the performance and adaptability of the sentiment classification model. Finally, the analysis phase synthesizes insights through detailed results analysis and documentation, providing a comprehensive understanding of the findings. This interconnected framework ensures that each phase contributes cohesively to the objectives, promoting methodological rigor and enabling adaptive improvements at every stage. Such a systematic approach is critical for addressing complex research problems and producing reliable, impactful outcomes tailored to specific domains.

The research framework systematically delineates the sequential processes essential for achieving the study's objectives, ensuring precision and methodological coherence. It commences with the data phase, which includes data collection, preprocessing, and exploratory analysis, to construct a solid foundation for subsequent modeling activities. Following this, the model phase emphasizes the sentiment classification model's iterative development, training, evaluation, and optimization, enhancing its accuracy and adaptability to intricate linguistic patterns. The final phase, analysis, integrates the outcomes through detailed results analysis and structured documentation, offering actionable insights and supporting the research's contribution to the academic field. This interconnected framework ensures alignment between all phases, fostering an adaptive approach to overcoming challenges and refining methodologies as necessary. By integrating these stages cohesively, the framework promotes a logical and systematic flow of activities, leading to reliable and impactful findings tailored to address complex research problems.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

2.2 Dataset

The dataset utilized in this study encompasses 5,796 customer reviews obtained from Agoda, offering a detailed representation of guest experiences and perceptions. Ratings are provided across multiple dimensions, including location (8.8), cleanliness (8.5), service quality (8.8), facilities (8.7), and room comfort (6.0), presenting both quantitative and qualitative evaluations. These ratings reflect diverse opinions and highlight aspects that meet or exceed expectations and areas requiring improvement, such as room comfort. The dataset's composition provides valuable insights into customer satisfaction, supported by specific comments and numerical scores that enrich the analysis. Such a dataset is particularly relevant for sentiment classification tasks as it combines structured and unstructured data, enabling a nuanced understanding of sentiment patterns. This comprehensive approach ensures that the dataset not only aids in sentiment analysis but also supports broader implications for improving service quality and enhancing guest experiences in the hospitality industry.

After eliminating incomplete, irrelevant, or poorly structured entries, the data-cleaning process led to the retention of 734 usable reviews for sentiment analysis. Initially, the dataset contained many reviews, but several were discarded due to missing ratings, inconsistent formatting, or unclear textual content. After a thorough cleaning, which involved filtering out invalid entries and ensuring the consistency of essential fields like ratings, review text, and sentiment labels, a refined subset of 734 reviews was obtained. These cleaned reviews provide a solid foundation for conducting sentiment analysis, as they represent the actual customer experience and sentiment, ensuring more reliable and meaningful insights. The result underscores the importance of data quality and preparation in sentiment analysis, emphasizing maintaining clean, structured data to achieve valid and actionable results.

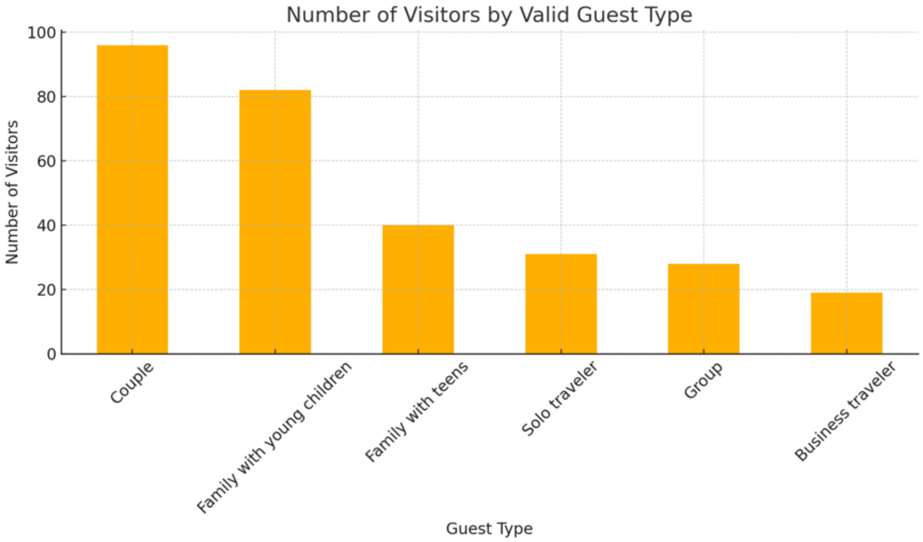


Figure 2. Number of Visitors by Valid Guest Type

Figure 2 illustrates the distribution of visitors categorized by valid guest types, providing detailed insights into customer demographics. The data reveals that "Couple" constitutes the largest segment, with approximately 100 visitors, followed closely by "Family with young children," which accounts for around 90 visitors. "Family with teens" and "Solo traveler" represent notable proportions, with roughly 60 and 50 visitors, respectively. Smaller segments include "Group," with approximately 40 visitors, and "Business traveler," with around 30 visitors. This distribution emphasizes the dominance of couples and families as primary customer groups, reflecting the hotel's strong appeal to these demographics. The comparatively lower numbers in the "Group" and "Business traveler" categories suggest untapped opportunities to attract these guest types. By strategically tailoring services and marketing initiatives to the preferences of these dominant segments while addressing the needs of underrepresented groups, the hotel can enhance guest satisfaction and expand its market reach effectively.

The distribution of visitors by valid guest types, expressed in percentages to provide a clearer understanding of customer composition. "Couple" emerges as the dominant category, accounting for approximately 28% of the total visitors, followed closely by "Family with young children" at 25%. "Family with teens" and "Solo traveler" represent 17% and 14% of visitors, respectively, showcasing their substantial contributions to the guest demographic. Smaller categories, such as "Group" and "Business traveler," constitute 11% and 8% of visitors, highlighting their relatively limited presence. These percentages emphasize the significant reliance on couples and families as the primary customer base, reflecting the hotel's strong alignment with these groups' preferences. The lower percentages of groups and business travelers suggest opportunities for growth through targeted marketing and tailored services. Strategically leveraging these insights allows the hotel to enhance its appeal to the dominant segments while addressing gaps to attract underrepresented groups, ensuring balanced growth and improved market positioning.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

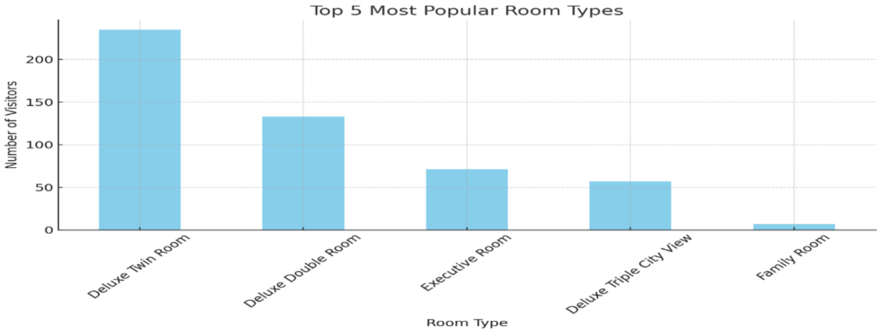


Figure 3. Top Five Most Popular Room Types

Figure 3 highlights the top five most popular room types based on visitor preferences, showcasing the distribution of bookings across these categories. The "Deluxe Twin Room" emerges as the most favored choice, with 235 bookings, accounting for a significant proportion of the total room selections. The "Deluxe Double Room" follows with 133 bookings, reflecting its strong appeal among guests seeking double occupancy. The "Executive Room" and "Deluxe Triple City View" hold third and fourth positions, with 71 and 57 bookings, respectively, suggesting a moderate preference for rooms with enhanced amenities or views. The "Family Room," with only seven bookings, indicates limited demand, likely due to its niche appeal for larger groups or families. This distribution reveals an apparent inclination towards twin and double rooms, likely driven by their versatility and affordability, making them suitable for various guest demographics. Such insights are instrumental in optimizing room allocation and marketing strategies to align with customer preferences, ensuring a balance between supply and demand for each room type.

The room data analysis demonstrates that the "Deluxe Twin Room" is the most preferred accommodation, comprising 45.5% of total bookings with 235 visitors, underscoring its popularity among diverse guest segments. The "Deluxe Double Room," capturing 25.7% with 133 visitors, emerges as the second most sought-after option, reflecting its strong appeal for couples and small groups. The "Executive Room," with 71 bookings (13.7%), and "Deluxe Triple City View," with 57 bookings (11%), indicate a moderate interest in the premium or view-oriented accommodations. Notably, the "Family Room," accommodating only seven visitors, reveals a niche preference and comparatively low demand. This distribution highlights the dominant role of twin and double room types in meeting the needs of most guests, likely due to their flexibility and cost-effectiveness. The insights emphasize the importance of prioritizing these room categories in inventory management and targeted marketing while exploring strategies to enhance the appeal of less popular options, ensuring a balanced and strategic approach to resource allocation.

3. RESULT AND DISCUSSION

3.1 IndoBert Model Performance Analysis

The dataset utilized in this study comprises textual reviews labeled into three sentiment categories: Negative, Neutral, and Positive. As illustrated in the "Dataset Label Distribution" graph, the dataset exhibits a significant class imbalance, with the Positive sentiment category overwhelmingly dominant, representing over 90% of the total samples. In contrast, the Negative and Neutral categories constitute only a tiny fraction of the dataset. This imbalance poses a critical challenge for the model, as it may develop a bias towards the majority class, potentially compromising its ability to identify minority classes accurately. Understanding this distribution is essential for addressing the inherent limitations in model training and evaluation, mainly when dealing with skewed datasets. Re-sampling or class-weight adjustments could mitigate these challenges and improve the model's performance across all sentiment categories.

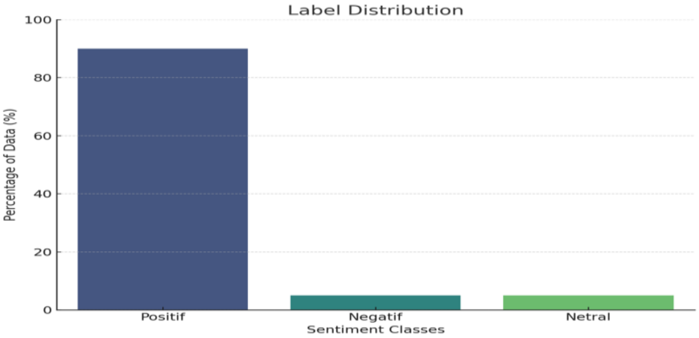


Figure 4. Label Distribution

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Figure 4 illustrates the distribution of sentiment labels in the dataset, providing insights into its composition and balance. The Positive class constitutes the majority, accounting for 90% of the total data, equivalent to 90 samples. In contrast, the Negative and Neutral classes comprise 5% of the dataset, corresponding to 5 samples per category. This disproportionate representation reveals a significant imbalance, which poses challenges for model training and evaluation. Such an imbalance may lead to a model that performs exceptionally well for the Positive class while neglecting the minority classes, resulting in biased predictions. Addressing this issue is critical, as it directly impacts the model’s ability to generalize effectively across all sentiment classes. Strategies such as re-sampling techniques, cost-sensitive learning, or data augmentation may be required to mitigate the effects of this imbalance and enhance the model's overall robustness. This analysis underscores the importance of balanced datasets in achieving reliable and equitable performance across categories.

The training history presents the trends in loss and accuracy over the epochs during the model’s training and validation phases. The training loss shows a steady decline in the loss graph, dropping from approximately 0.35 in the first epoch to below 0.15 in the third, indicating that the model is learning effectively from the training data. However, the validation loss follows a less consistent trajectory, initially decreasing but showing an increase in the final epoch, which suggests that the model may be beginning to overfit the training data. The training accuracy improves consistently, rising from around 90% in the first epoch to over 95% in the final epoch, demonstrating the model's growing capability to classify the training data correctly. On the other hand, the validation accuracy remains relatively stable, fluctuating slightly around 92%, with a minor dip in the second epoch before a slight recovery. The observed gap between training and validation accuracy, coupled with the increase in validation loss, highlights a potential overfitting issue, where the model is overly specialized to the training data at the expense of generalization to unseen data. These trends emphasize the need for careful evaluation and mitigation strategies to address overfitting, such as implementing regularization techniques, utilizing early stopping, or collecting more balanced data to improve the model's generalizability across all sentiment classes.

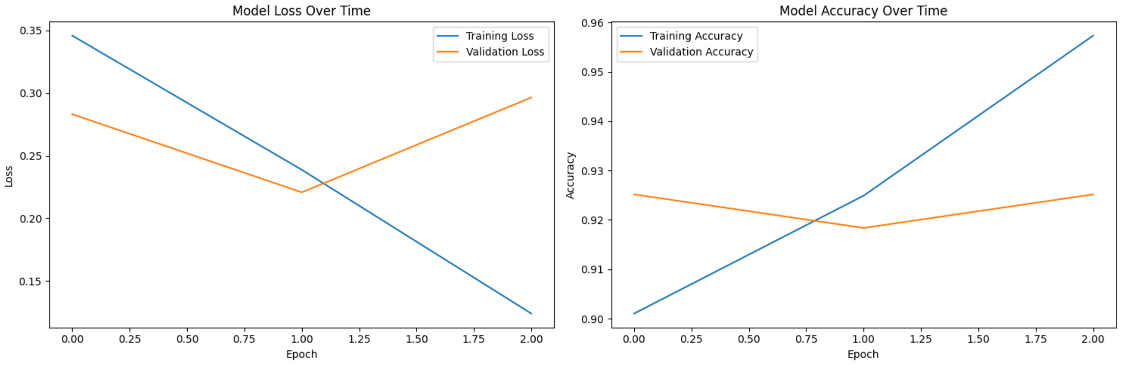


Figure 5. Training History

Figure 5 illustrates the model's training history, highlighting the trends in loss and accuracy over multiple epochs for both the training and validation phases. The training loss exhibits a consistent decline, dropping from approximately 0.35 in the first epoch to below 0.15 by the final epoch, indicating effective learning from the training data. Conversely, the validation loss initially decreases but shows an upward trend after the first epoch, suggesting potential overfitting as the model becomes overly specialized to the training data. Similarly, the training accuracy steadily improves, surpassing 95% by the final epoch, while the validation accuracy stabilizes around 92%, with a minor fluctuation in the intermediate epoch. This gap between training and validation metrics underscores the challenges of generalizing unseen data. Such observations emphasize the importance of regularization techniques and balanced datasets to mitigate overfitting and enhance model robustness for broader applications.

The classification report provides an in-depth evaluation of the model's performance, highlighting global accuracy and detailed metrics for each sentiment class. The model's overall accuracy is 92.52%, indicating strong performance in correctly classifying most of the data. However, a closer examination of per-class metrics reveals significant disparities in performance. For the Negative class, the model achieves a precision of 1.00, suggesting that all predictions made for this class are correct. However, the recall is extremely low at 0.1667, indicating that the model identifies only a tiny fraction of actual Negative instances. This results in an f1-score of just 28.57%, underscoring the model's inability to generalize effectively for this minority class. The Neutral class performs even worse, with precision, recall, and f1-score all at 0.00, signifying that the model fails to detect this class completely. This is likely due to the highly imbalanced dataset, where Neutral samples constitute a tiny proportion of the data. In contrast, the Positive class exhibits robust performance, with a precision of 0.9310 and a recall of 0.9926, resulting in a high f1-score of 96.09%. This reflects the model's strong ability to classify the dominant class correctly, which aligns with the high representation of this class in the dataset. The macro-averaged scores, which treat all classes equally, reveal significant weaknesses, with precision at 0.6437, recall at 0.3864, and f1-score at

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 976−986
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6505

0.4155. These metrics highlight the model's struggles with minority classes despite its success with the Positive class. Addressing these performance gaps may require re-sampling techniques, adjusting class weights, or incorporating additional data to enhance the representation of the underperforming classes.

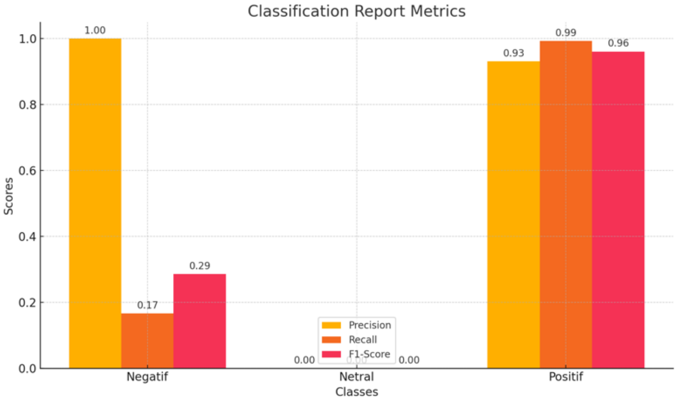


Figure 6. Classification Report Metrics

Figure 6 illustrates the classification report metrics, providing a detailed evaluation of the model's performance across the sentiment classes. The Positive class demonstrates consistently strong metrics, with a precision of 0.93, a recall of 0.99, and an F1-score of 0.96, indicating the model's ability to accurately identify and classify the dominant class. In contrast, the Negative class achieves a perfect precision of 1.00. Still, its recall is significantly lower at 0.17, leading to an F1-score of only 0.29, reflecting the model's inability to generalize for this minority class effectively. The Neutral class, however, performs the weakest, with precision, recall, and F1-score all at 0.00, highlighting a complete failure to predict this class. This disparity underscores the challenges posed by the dataset's class imbalance, which skews the model's focus towards the majority class at the expense of minority classes. Addressing this issue is crucial to ensure a balanced and fair classification model, and strategies such as class rebalancing or advanced loss functions could be explored to improve performance across all classes.

The confusion matrix and normalized confusion matrix provide a comprehensive view of the model's predictions across the sentiment classes. The Positive class dominates the predictions, with most labels correctly classified (99.26%). However, this overwhelming focus on the Positive class reveals a notable bias in the model's performance. The Negative class, despite achieving a high precision of 1.00, is underrepresented in the recall, as only 16.67% of actual Negative labels are correctly predicted, while the remaining are misclassified as Positive. The Neutral class performs the poorest, with none of its actual labels being correctly predicted, as all instances are misclassified into other classes, particularly Positive. This stark imbalance highlights the model's reliance on patterns heavily represented in the Positive class, a direct consequence of the dataset's class imbalance. The normalized confusion matrix further emphasizes these trends, showing that the model disproportionately favors the Positive class in its predictions. Addressing this challenge requires strategies such as rebalancing the dataset, applying advanced sampling techniques, or modifying the loss function to penalize misclassifications in minority classes more heavily. These interventions would enhance the model's ability to generalize effectively across all sentiment classes and mitigate the current bias.

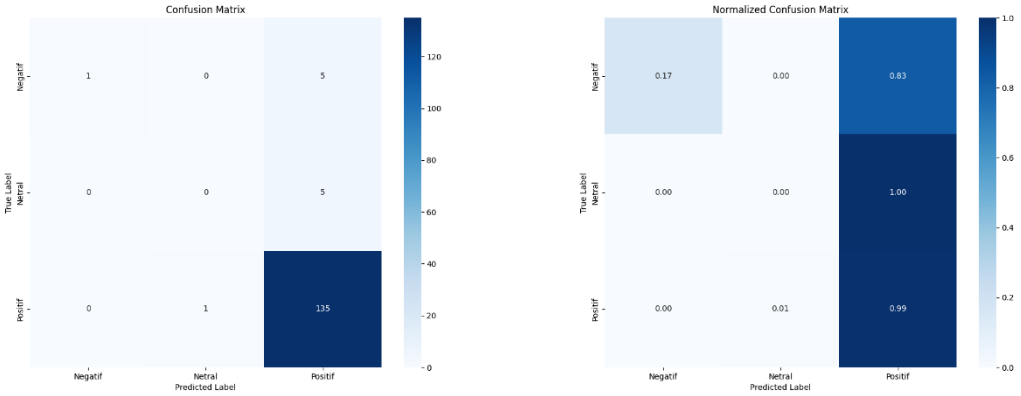
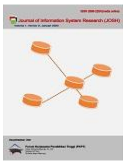


Figure 7. Confusion Matrix and Normalized Confusion Matrix

Figure 7 illustrates the confusion matrix and its normalized counterpart, offering a detailed evaluation of the model's prediction distribution across sentiment classes. The Positive class exhibits strong performance, with



99.26% of actual Positive instances accurately predicted, demonstrating the model's proficiency in identifying this dominant class. However, the Negative class presents significant limitations; despite achieving perfect precision, only 16.67% of actual Negative samples are correctly classified, with the majority misclassified as Positive. The Neutral class performs the weakest, with no actual Neutral samples being correctly identified, as all predictions are erroneously allocated to other classes, predominantly Positive. This imbalance in performance highlights the model's bias towards the Positive class, likely resulting from the dataset's skewed class distribution. The normalized confusion matrix further emphasizes this trend, showing the disproportionately high prediction rate for the Positive class at the expense of the minority classes. Such a bias limits the model's generalization capabilities, particularly in accurately identifying underrepresented categories. Addressing this issue necessitates strategies such as balanced data augmentation, reweighting loss functions, or implementing ensemble methods to enhance performance across all classes while reducing over-reliance on the dominant class. These adjustments are critical to developing a more equitable and robust sentiment classification model.

The Precision-Recall curves, as illustrated in the figure, provide insights into the model's performance in handling imbalanced data. The Positive class exhibits the highest Average Precision (AP) score of 0.99, indicating the model's intense focus and ability to predict this dominant class accurately. This high AP score reflects the model's capability to balance precision and recall for Positive predictions, even when thresholds are adjusted. In contrast, the Negative class demonstrates a significantly lower AP score of 0.42, highlighting the model's struggles with identifying actual Negative instances amidst the overwhelming prevalence of the Positive class. The Neutral class performs the poorest, with an AP score of 0.12, signifying its near-total absence in accurate predictions due to its sparse representation in the dataset. These disparities underscore the model's bias towards the Positive class, a direct consequence of the imbalanced class distribution in the dataset. While the high AP for the Positive class confirms the model's strength in predicting the dominant category, the drastically lower AP scores for Negative and Neutral classes indicate a need for targeted interventions. Strategies such as rebalancing the dataset, introducing synthetic samples for minority classes, or utilizing class-aware loss functions may help improve the model's generalization across all classes, ensuring a more equitable performance.

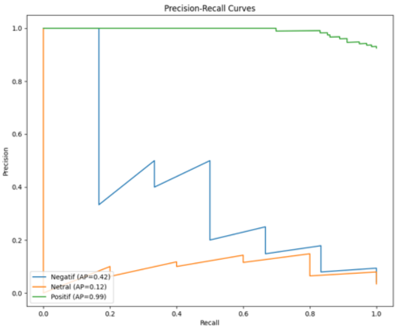


Figure 8. Precision-Recall Curves

Figure 8 presents the Precision-Recall curves, which evaluate the model's performance in managing precision and recall across varying thresholds for each sentiment class. The Positive class demonstrates exceptional performance, with an Average Precision (AP) score of 0.99, reflecting the model's strong ability to maintain high precision and recall when predicting this dominant class. This outcome is consistent with the dataset's significant imbalance, where the Positive class constitutes the majority of samples. Conversely, the Negative class shows a markedly lower AP score of 0.42, indicating that the model struggles to achieve a reliable balance between precision and recall for this minority class. The Neutral class performs the weakest, with an AP score of only 0.12, signifying the model's near inability to distinguish and predict instances of this class correctly. The poor performance of these minority classes underscores the challenges the highly skewed data distribution poses. This analysis highlights the necessity for strategies to address class imbalance, such as applying re-sampling techniques, adjusting class weights during training, or utilizing advanced architectures tailored to imbalanced datasets. These interventions are critical to improving the model's generalization ability across all classes and ensuring more equitable performance. The high AP for the Positive class should be leveraged while enhancing the reliability of predictions for the underrepresented categories.

As depicted in the figure, the ROC curves represent the model's ability to differentiate between classes. The area under the curve (AUC) quantitatively measures the model's classification performance. The Positive class achieves the highest AUC of 0.90, reflecting the model's strong capability to distinguish this majority class from the others. This high AUC underscores the model's focus on the Positive class, which aligns with its dominant representation in the dataset. The Negative class exhibits an AUC of 0.88, indicating reasonably good performance despite being a minority class. However, this result is likely influenced by the model's high precision but limited recall for this class, as seen in other metrics. On the other hand, the Neutral class has the lowest AUC at 0.83, which highlights the model's struggle to accurately identify and separate instances of this class. This performance

This Journal is licensed under a Creative Commons Attribution 4.0 International License

gap reflects the imbalanced dataset's challenges, where the neutral class is underrepresented, making it more difficult for the model to predict effectively. These observations emphasize the importance of addressing dataset imbalance and optimizing the model's architecture to enhance its generalization ability across all classes. While the high AUC for the Positive class showcases the model's strength in handling the dominant class, improving AUC scores for the minority classes is critical for achieving a more balanced and equitable classification system. Techniques such as cost-sensitive learning, re-sampling, or ensemble methods could be explored to address these challenges.

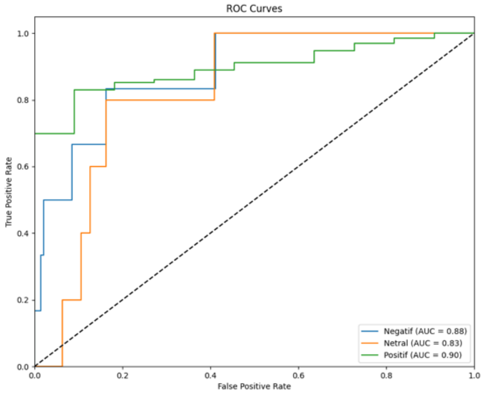


Figure 9. ROC Curves

Figure 9 depicts the ROC curves, illustrating the model's ability to distinguish between true positive and false positive rates for each sentiment class across varying decision thresholds. The Positive class achieves the highest area under the curve (AUC) at 0.90, signifying the model's superior capability in correctly identifying instances of this majority class. This result reflects the model's alignment with the dataset's inherent bias, as the Positive class constitutes the dominant representation of the data. The Negative class demonstrates a relatively high AUC of 0.88, indicating that the model exhibits a good balance between sensitivity and specificity for this class, albeit limited by the dataset's imbalance. The Neutral class, however, achieves the lowest AUC at 0.83, underscoring the model's difficulty in accurately predicting instances from this underrepresented category. The weaker performance for this class highlights the challenges associated with imbalanced datasets, where minority classes often suffer from poor generalization due to inadequate representation during training. These findings emphasize the importance of implementing strategies to improve model performance for minority classes without compromising the accuracy of the majority class. Techniques such as class rebalancing, synthetic data generation, and adaptive loss functions enhance the model's capacity to generalize across all classes, ensuring more equitable and robust classification outcomes. The high AUC values for the Positive and Negative classes provide a foundation for further refinements to address the disparities identified in the Neutral class.

The evaluation highlights the model's strong global performance, with an overall accuracy of 92.52% and a high AUC for the Positive class (0.90), reflecting its ability to effectively classify the dominant sentiment category. However, significant weaknesses are evident in the model's handling of minority classes, as demonstrated by the low F1 scores for the Negative class (28.57%) and the inability to detect the Neutral class. These disparities arise from the pronounced class imbalance in the dataset, which skews the model's predictions heavily toward the Positive class. To address these issues, future research should prioritize techniques to mitigate data imbalance. Oversampling methods, such as SMOTE, could increase the representation of minority classes, thereby enhancing the model's exposure to diverse patterns during training. Additionally, reweighting the loss function to penalize misclassifications in underrepresented classes more heavily would encourage the model to allocate more significant attention to these categories. Incorporating ensemble methods or hybrid architectures may further strengthen the model's robustness and generalization across all sentiment classes. Combined with a broader and more balanced dataset, these strategies would support the development of a more equitable and reliable sentiment classification framework.

4. CONCLUSION

This research focuses on analyzing the performance of a sentiment classification model using IndoBERT on Indonesian hotel review data. Sentiment analysis plays a critical role in understanding customer feedback, especially in the hospitality industry, where insights from reviews influence service improvement and decision-making. However, challenges such as linguistic complexity and imbalanced datasets often hinder the performance of sentiment classification models. This study leverages IndoBERT, a transformer-based pre-trained language model, due to its proven capability to understand Indonesian text. The primary objective is to evaluate the model's

This Journal is licensed under a Creative Commons Attribution 4.0 International License

effectiveness in classifying sentiment into Positive, Neutral, and Negative categories and identify improvement areas, particularly in handling imbalanced datasets. The research employed a three-phase methodology: data preprocessing, model training, and evaluation. The dataset included textual hotel reviews labeled into three sentiment categories: Positive, Neutral, and Negative, with the Positive class dominating at 90%, while Neutral and Negative classes constituted 5% each. During preprocessing, reviews were tokenized and transformed using IndoBERT's tokenizer. The dataset was split into training (80%) and validation (20%) sets. A custom PyTorch dataset class was created to manage inputs and labels. The IndoBERT model, fine-tuned for sequence classification with three output labels, was trained over three epochs using AdamW as the optimizer and cross-entropy loss. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC and Precision-Recall curves. The results demonstrated high global accuracy (92.52%) and excellent performance for the Positive class, with an F1 score of 96.09%. However, the model struggled with minority classes, particularly Neutral, where the precision, recall, and F1-score were all 0.00. These findings highlight the need for strategies to address data imbalance, such as oversampling or reweighting loss functions, to improve performance for underrepresented classes.

ACKNOWLEDGMENT

The authors sincerely thank the Atma Jaya Catholic University of Indonesia, particularly the Faculty of Business Administration and Communication and Tourism Study Program, for their institutional support in conducting this research. Special thanks to LPPM (Institute for Research and Community Service) and PUSDIPAR (Center of Digital Transformation and Tourism Development) for providing research facilities and guidance throughout this study.

REFERENCES

[1] D. K. Kardaras, C. Troussas, S. G. Barbounaki, P. Tselenti, and K. Armyras, "A Fuzzy Synthetic Evaluation Approach to Assess Usefulness of Tourism Reviews by Considering Bias Identified in Sentiments and Articulacy," *Inf.*, vol. 15, no. 4, 2024, doi: 10.3390/info15040236.

[2] R. A. Rahman and Suyanto, "Performance Analysis of ChatGPT for Indonesian Abstractive Text Summarization," in *Proceedings - International Seminar on Intelligent Technology and its Applications, ISITIA*, 2024, no. 2024, pp. 477–482. doi: 10.1109/ISITIA63062.2024.10668361.

[3] M. T. Uliniansyah *et al.*, "Twitter dataset on public sentiments towards biodiversity policy in Indonesia," *Data Br.*, vol. 52, 2024, doi: 10.1016/j.dib.2023.109890.

[4] K. Purwandari, M. A. Jiwanggi, and E. Yulianti, "Sentiment Analysis on YouTube Comment Data for the Candidate Debate in the 2024 Presidential Election of the Republic of Indonesia," in *2024 5th International Conference on Artificial Intelligence and Data Sciences, AiDAS 2024 - Proceedings*, 2024, pp. 392–397. doi: 10.1109/AiDAS63860.2024.10730443.

[5] I. Daqiqil, H. Saputra, Syamsudhuha, R. Kurniawan, and Y. Andriyani, "Sentiment analysis of student evaluation feedback using transformer-based language models," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 36, no. 2, pp. 1127–1139, 2024, doi: 10.11591/ijeecs.v36.i2.pp1127-1139.

[6] M. Irdayanti, D. Purwitasari, and D. O. Siahaan, "Relevance Detection using Text Entailment for Health-related Question-Answer Texts with Imbalanced Data," in *Proceedings - International Seminar on Intelligent Technology and its Applications, ISITIA*, 2024, no. 2024, pp. 681–686. doi: 10.1109/ISITIA63062.2024.10667778.

[7] E. Yulianti, N. Bhary, J. Abdurrohman, F. W. Dwitilas, E. Q. Nuranti, and H. S. Husin, "Named entity recognition on Indonesian legal documents: a dataset and study using transformer-based models," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 5, pp. 5489–5501, 2024, doi: 10.11591/ijece.v14i5.pp5489-5501.

[8] G. Enrique, I. Alfina, and E. Yulianti, "Javanese part-of-speech tagging using cross-lingual transfer learning," *IAES Int. J. Artif. Intell.*, vol. 13, no. 3, pp. 3498–3509, 2024, doi: 10.11591/ijai.v13.i3.pp3498-3509.

[9] E. I. Setiawan *et al.*, "Indonesian News Stance Classification Based on Hybrid Bidirectional LSTM and Transformer Based Embedding," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, pp. 517–537, 2024, doi: 10.22266/ijies2024.1031.41.

[10] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Recognizing Indonesian Textual Entailment," in *Proceedings - International Conference on Informatics and Computational Sciences*, 2024, pp. 462–467. doi: 10.1109/ICICoS62600.2024.10636863.

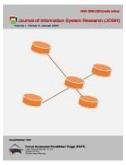
[11] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, 2024, doi: 10.1155/2024/2826773.

[12] 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 Retrieval," in *KST 2024 - 16th International Conference on Knowledge and Smart Technology*, 2024, pp. 242–247. doi: 10.1109/KST61284.2024.10499674.

[13] Y. A. A. I. Rifai and D. Suhartono, "Emotion Classification of Indonesian Twitter Social Media Text Using Soft Voting Ensemble Method," *ICIC Express Lett. Part B Appl.*, vol. 15, no. 1, pp. 101–108, 2024, doi: 10.24507/icicelb.15.01.101.

[14] Edwina and T. Mauritsius, "Data-Driven Insights for Mobile Banking App Improvement: A Sentiment Analysis and Topic Modelling Approach for SimobiPlus User Reviews," *Int. J. Eng. Trends Technol.*, vol. 72, no. 6, pp. 347–360, 2024, doi: 10.14445/22315381/IJETT-V72I6P132.

[15] S. Latisha, S. Favian, and D. Suhartono, "Criminal Court Judgment Prediction System Built on Modified BERT Models," *J. Adv. Inf. Technol.*, vol. 15, no. 2, pp. 288–298, 2024, doi: 10.12720/jait.15.2.288-298.



[16] F. S. Yerzi, D. P. Ramadhani, and A. Alamsyah, "Comparison of Multiclass Classification and Topic Modeling to Identify Technology Acceptance in Popular E-Commerce in Indonesia Based on UTAUT3 Model," in *Proceedings of the 2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2024*, 2024, pp. 273–279. doi: 10.1109/IAICT62357.2024.10617771.

[17] J. Islamey, V. Jonathan, M. Nurzaki, and H. Lucky, "Comparative Analysis of Encoder-Based Pretrained Models: Investigating the Performance of BERT Variants in Indonesian Question-Answering," in *2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics, ICoABCD 2024*, 2024, pp. 309–314. doi: 10.1109/ICoABCD63526.2024.10704260.

[18] A. F. Hidayatullah, "Code-Mixed Sentiment Analysis on Indonesian-Javanese-English Text Using Transformer Models," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2024*, 2024, pp. 340–345. doi: 10.1109/ICITISEE63424.2024.10730138.

[19] R. A. Fitrianto, A. S. Editya, M. M. H. Alamin, A. L. Pramana, and A. K. Alhaq, "Classification of Indonesian Sarcasm Tweets on X Platform Using Deep Learning," in *Proceedings - International Conference on Informatics and Computational Sciences*, 2024, pp. 388–393. doi: 10.1109/ICICoS62600.2024.10636904.

[20] R. Sivanaiah, S. Suresh, S. Pandian, and A. D. Suseelan, "Bridging the Language Gap: Transformer-Based BERT for Fake News Detection in Low-Resource Settings," *Communications in Computer and Information Science*, vol. 2046 CCIS. pp. 398–411, 2024. doi: 10.1007/978-3-031-58495-4_29.

[21] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.

[22] F. V. P. Samosir and S. Riyaldi, "Sentiment Analysis of TikTok Comments on Indonesian Presidential Elections Using IndoBERT," in *2024 3rd International Conference on Creative Communication and Innovative Technology, ICCIT 2024*, 2024. doi: 10.1109/ICCIT62134.2024.10701256.

[23] H. M. Ramdhan, M. Dwifebri Purbolaksono, and B. Bunyamin, "Sentiment Analysis of Beauty Product Reviews Using the IndoBERT Method and Naive Bayes Classification," in *2024 12th International Conference on Information and Communication Technology, ICoICT 2024*, 2024, pp. 397–404. doi: 10.1109/ICoICT61617.2024.10698198.

[24] K. Chandra, K. A. Prasetya, R. D. Saputra, and M. F. Hasani, "Leveraging IndoBert for CyberBullying Classification on Social Media," in *ICSINTESA 2024 - 2024 4th International Conference of Science and Information Technology in Smart Administration: The Collaboration of Smart Technology and Good Governance for Sustainable Development Goals*, 2024, pp. 407–411. doi: 10.1109/ICSINTESA62455.2024.10747874.

[25] E. Dave and A. Chowanda, "IPerFEX-2023: Indonesian personal financial entity extraction using indoBERT-BiGRU-CRF model," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00987-6.

[26] K. E. Saputra and Riccosan, "Indonesian news article authorship attribution multilabel multiclass classification using IndoBERT," *IAES Int. J. Artif. Intell.*, vol. 13, no. 4, pp. 4688–4694, 2024, doi: 10.11591/ijai.v13.i4.pp4688-4694.

[27] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[28] F. Rahman and A. S. Girsang, "IndoBERTweet for Sarcasm: Evaluating Domain-Adapted Transformers for Indonesian Twitter Sarcasm Classification," *J. Logist. Informatics Serv. Sci.*, vol. 11, no. 2, pp. 155–164, 2024, doi: 10.33168/JLISS.2024.0210.

[29] H. Santosa, F. Rachman, S. A. Austen, Christianto, and A. S. Girsang, "IndoBERT for classifying hate speech in Twitter," in *AIP Conference Proceedings*, 2024, vol. 3026, no. 1. doi: 10.1063/5.0199750.

[30] R. N. Tanaja, A. Widjaya, Johnny, A. A. S. Gunawan, and K. E. Setiawan, "Evaluating Public Opinion on the 2024 Indonesian Presidential Election Candidate: An IndoBERT Approach to Twitter Sentiment Analysis," in *2024 10th International Conference on Smart Computing and Communication, ICSCC 2024*, 2024, pp. 88–94. doi: 10.1109/ICSCC62041.2024.10690796.