Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 1, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i1.10021

Deep Learning Sentiment Analysis of Islamic Boarding School Google Reviews Using IndoBERT Variants and XLM-RoBERTa

Ahmad Syarifuddin[1], Endang Wahyu Pamungkas[1], Muhammad Muharrom Al Haromainy[2]
[1]Univeristas Muhammadiyah Surakarta, Indonesia
[2]Universitas Pembangunan Nusantara Veteran Jawa Timur, Indonesia
asyariiif@gmail.com, ewp123@ums.ac.id, muhammad.muharrom.if@upnjatim.ac.id

Info Artikel

Riwayat Artikel:

Received 2025-12-24
Revised 2026-02-10
Accepted 2026-02-20

Abstract – *Online reviews on platforms like Google Maps have become a crucial data source for analyzing public opinion and institutional reputation, including for Islamic boarding schools (pesantren). Negative stigmas arising from recent incidents have made understanding public perception vital for institutional evaluation. This study aims to perform sentiment analysis to measure public perception of pesantren across Java using state-of-the-art deep learning models. A dataset of 8,577 reviews was collected via web scraping and underwent comprehensive preprocessing, including cleansing, case folding, tokenization, stopword removal, and stemming. The data were partitioned using a Stratified Train-Test Split (70:30). This research evaluates and compares three pre-trained language models—IndoBERT Base, IndoROBERTa Small, and XLM-RoBERTa—fine-tuned with a Focal Loss function to address significant class imbalances. The final evaluation demonstrated that the IndoBERT Base model significantly outperformed the others, achieving an overall accuracy of 92%. The model showed balanced performance across all sentiment categories, effectively mitigating classification bias through the Focal Loss mechanism IndoBERT Base is identified as the optimal model for this domain. The study provides a robust framework for religious-educational institutions to monitor their digital reputation and gain actionable feedback from public reviews.*

Keywords: *IndoBERT, IndoROBERTa, Pesantren, Sentiment Analysis, XLM-RoBERTa*

Corresponding Author:
Endang Wahyu Pamungkas
Email: ewp123@ums.ac.id

Abstrak – *Ulasan daring pada platform seperti Google Maps telah menjadi sumber data penting untuk menganalisis opini publik dan reputasi institusi, termasuk bagi pesantren. Stigma negatif yang muncul dari berbagai insiden baru-baru ini menjadikan pemahaman terhadap persepsi masyarakat sangat krusial untuk evaluasi kelembagaan. Penelitian ini bertujuan untuk melakukan analisis sentimen guna mengukur persepsi masyarakat terhadap pesantren di seluruh Pulau Jawa menggunakan model deep learning terkini. Dataset sebanyak 8.577 ulasan dikumpulkan melalui teknik web scraping dan melalui tahap preprocessing lengkap yang meliputi cleansing, case folding, tokenization, penghapusan stopword, dan stemming. Data dibagi menggunakan metode Stratified Train-Test Split (70:30). Penelitian ini mengevaluasi dan membandingkan tiga model bahasa pralatih—IndoBERT Base, IndoROBERTa Small, dan XLM-RoBERTa—yang di-fine-tuning menggunakan fungsi Focal Loss untuk menangani ketidakseimbangan kelas yang signifikan. Evaluasi akhir menunjukkan bahwa model IndoBERT Base mengungguli model lainnya secara signifikan dengan tingkat akurasi keseluruhan sebesar 92%. Model ini menunjukkan kinerja yang seimbang di seluruh kategori sentimen dan secara efektif mengurangi bias klasifikasi melalui mekanisme Focal Loss. IndoBERT Base diidentifikasi sebagai model paling optimal untuk domain ini. Penelitian ini memberikan kerangka kerja yang kuat bagi institusi pendidikan keagamaan untuk memantau reputasi digital mereka dan memperoleh umpan balik yang dapat ditindaklanjuti dari ulasan publik.*

Kata Kunci*: IndoBERT, IndoROBERTa, Pesantren, Sentiment Analysis, XLM-RoBERTa*

I. INTRODUCTION

Islamic boarding schools (*pesantren*) are among Indonesia's oldest educational institutions, playing a vital role in character formation and religious development [1]. As they evolve, *pesantren* have integrated general subjects into their curriculum to meet both spiritual and worldly needs [2]. However, frequent negative news coverage, such as incidents of violence and misconduct, has damaged their reputation and created a negative stigma among the public [3]. Understanding public perception in the wake of such incidents is now crucial for *pesantren* administrators and stakeholders to evaluate their institutional image.

In the digital era, Google Maps has emerged as a primary platform for the public to share experiences and evaluate services [4]. Ratings and reviews on this platform offer significant benefits, providing accessibility and familiarity that influence buying and selection decisions [5]. Specifically for education, many guardians now rely on Google Maps reviews as a primary guide for choosing a *pesantren*, where these evaluations impact the institution's reputation, sustainability, and social value [6].

Sentiment analysis offers a systematic way to study these public opinions and emotions [7]. By applying text mining to review columns, researchers can determine end-user perceptions regarding the quality of services provided [8]. Previous studies, such as the analysis of the Alfagift application, demonstrate that sentiment analysis is a strategic tool for identifying institutional strengths and weaknesses to drive data-driven improvements [9]. In an educational context, it serves as a feedback mechanism for learning effectiveness and curriculum development [10].

Several studies have explored different methods for sentiment classification. Research on the Maxim application utilized Long Short-Term Memory (LSTM) to capture sequential patterns, though it suggested the use of BERT for better performance model [11]. In the Indonesian context, IndoBERT has shown outstanding results in dirty vote film documenter, achieving up to 99% accuracy [12]. Meanwhile, IndoROBERTa has demonstrated robustness in processing Indonesian news, despite challenges with class imbalance [13]. Furthermore, multilingual models like XLM-RoBERTa have proven powerful in low-resource settings and cross-lingual generalization, outperforming traditional models [14].

However, the complexity of Indonesian user-generated content—often containing non-standard language and religious code-mixing—requires optimal model selection [15]. While specialized models like IndoBERT and IndoROBERTa are optimized for Indonesian structures, multilingual models like XLM-RoBERTa offer different capacities for generalization [16]. Therefore, this study aims to develop and compare a deep learning-based sentiment analysis model using IndoBERT base, IndoROBERTa small, and XLM-RoBERTa base [17].

The novelty of this research lies in its specific focus on the *pesantren* educational domain using a large-scale, randomly sampled dataset from Google Maps reviews across Java. Unlike previous studies that often focus on commercial applications or general news, this research addresses the unique linguistic nuances and class imbalances inherent in religious educational reviews. Furthermore, this study explicitly contributes by implementing and evaluating the Focal Loss function across three state-of-the-art Pre-trained Language Models (PLMs) to handle sentiment disparity. This comparison establishes the most robust framework for capturing public perception in a sensitive educational context, providing a strategic evaluation tool that has not been extensively explored in prior *pesantren*-related literature.

II. METHODS

Methods The systematic workflow of this research is illustrated in Figure 1, encompassing stages from data collection to final model evaluation.

A. Data Collection and Sampling

The data utilized in this study were collected from the Google Maps platform using a web scraping tool designed to efficiently extract review data, including the text content and associated ratings. To ensure a representative sample for this educational context, the focus was placed specifically on Islamic boarding schools (pesantren) located across the island of Java, with the selection process being carried out randomly to maximize geographical and institutional diversity. This systematic collection process yielded a substantial dataset of 8,577 total reviews. Subsequently, the collected data were structured and saved in a Comma Separated Values (CSV) file format to facilitate easy storage, portability, and streamlined processing by analytical software and programming scripts [18].

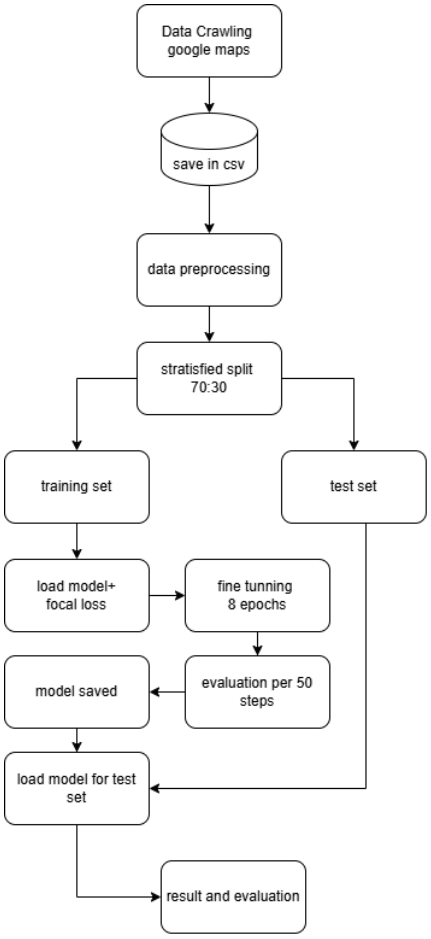

Figure 1. Methods used in this research

B. Data Preparation and Labeling

Following the scraping process, the raw data underwent a crucial manual labeling stage to establish a ground truth for sentiment analysis [19]. A subset of the 8,577 reviews was carefully labeled: 1,165 reviews were classified as having a negative sentiment, 1,790 reviews were labeled as neutral, and 4,000 reviews were identified as expressing a positive sentiment. The remaining portion of the dataset was reserved for model evaluation and testing purposes. This carefully labeled dataset ensures the robustness and accuracy of the subsequent training and testing phases of the sentiment analysis model, allowing the research to accurately ascertain public perception regarding the surveyed pesantren.

After the data labeling process, word frequencies were analyzed to identify prominent terms in each sentiment category, as visualized through the word clouds in Figure 2 (Positive), Figure 3 (Negative), and Figure 4 (Neutral). These frequencies are then presented visually through word clouds. Each review is associated with a different set of words, whether categorized as positive, negative or neutral as shown in these Figure.

Figure 2. Positive Word Cloud

Figure 3. Negative Word Cloud

Figure 4. Netral Word Cloud

Before the application of linguistic cleaning techniques (such as case folding, cleansing, tokenization, stopword removal, and stemming), the dataset underwent essential preliminary data preprocessing. This stage began with a cleansing phase to address structural issues, specifically the removal of any instances of missing values (NaN) and unnecessary whitespace [20]. Following this, a crucial step of label standardization was performed to ensure consistency across the sentiment categories: the misspelled label "nertral" was corrected to "netral," and all English sentiment labels ("positive," "negative") were harmonized to their Indonesian equivalents ("positif" and "negatif," respectively). Finally, these standardized categorical labels were converted into a numerical format for machine learning compatibility, where "negatif" was mapped to 0, "netral" to 1, and "positif" to 2.

C. Case Folding

Case folding is a fundamental text preprocessing step that involves converting all textual data to lowercase [21], strictly retaining only alphabetic characters from 'a' to 'z'. Any characters that are not letters—such as punctuation marks, numbers, or symbols—are treated as delimiters and consequently removed from the text [22]. For instance, the original phrase, ‘Pendidikan Pondok PESAntren itu LUAR BIASA!’ (The education at the Islamic boarding school is EXTRAORDINARY!), is simplified through case folding to: ‘pendidikan pondok pesantren itu luar biasa’ (the education at the Islamic boarding school is extraordinary). This standardization ensures uniformity and prepares the text for subsequent analysis.

D. Cleansing and Tokenization

Cleansing is the process of removing all unnecessary or irrelevant elements from the text [23]. This crucial step typically involves eliminating items that do not contribute to the meaning or sentiment, such as URLs, numbers, symbols, and irrelevant punctuation marks [24]. This ensures that only meaningful text remains for analysis. For example, the noisy original text: ‘Wow!! This Islamic boarding school is so cool! 😎👍, or call 08123456789 #ayoMondok!!!’, is cleaned to a simplified and readable form: ‘Wow, this Islamic boarding school is so cool’[25].

Tokenization is the process of dissecting a continuous stream of text or a document into smaller, discrete components known as tokens, which are typically individual words or phrases [26]. This is an essential step that breaks down the text into manageable units for linguistic analysis. For instance, the original sentence: ‘boarding school education is extraordinary!’, is transformed through tokenization into a list of individual tokens: [‘education’, ‘boarding school’, ‘it’, “extraordinary”, ‘!’].

E. Stopword Removal and Stemming

Stopword removal is a critical filtering process aimed at eliminating words that appear frequently but carry minimal semantic or emotional value in the context of sentiment analysis [27]. These are common connecting words, often referred to as "noise", which can unnecessarily inflate the dimensionality of the data without improving the model’s performance. Examples of stopwords in Indonesian include 'yang', 'di', and 'itu'. By removing these terms,

the focus is sharpened onto the words that truly express opinion. Before removal: ‘pendidikan’, ‘di’, ‘pesantren’, ‘itu’, ‘sangat’, ‘baik’, and after removal: ‘pendidikan’, ‘pesantren’, ‘sangat’, ‘baik’

Stemming is a normalization technique designed to transform all word variations back to their root or base form [28]. This process is vital for ensuring that words with the same core meaning (e.g., different suffixes or prefixes) are treated as a single token during analysis, thereby reducing data redundancy. For example, in the Indonesian language, the words 'berkunjung', 'mengunjungi', and 'kunjungan' all share the base word 'kunjung'. This unification is essential for accurate frequency counting and analysis. Before stemming: ‘pendidikan’, ‘mengajarkan’, ‘kedamaian’, and after stemming: ‘didik’, ‘ajar’, ‘damai’.

F. Stratified Train-Test Split

To ensure that the distribution of sentiment categories (Positive, Negative, and Neutral) remains consistent between the training and testing sets, the data must be partitioned using the Stratified Train-Test Split method [29]. This technique is crucial, especially when dealing with imbalanced datasets, as it maintains the proportional representation of each class across both subsets [30], thereby preventing training bias and ensuring the model is evaluated on a truly representative sample. In this study, the entire labeled dataset was divided into two distinct subsets: 70% of the data was allocated for the training set to optimize the model parameters, and the remaining 30% of the data was reserved for the testing set to assess the final model performance and generalization capability.

G. Load Pre-trained Model

For the sentiment classification task, the research utilized several powerful pre-trained Language Models (LMs) to leverage existing linguistic knowledge. The specific models chosen for evaluation include IndoBERT (specifically indobenchmark/IndoBERT -base-p1), IndoROBERTa Small, and XLM-RoBERTa. Correspondingly, the respective IndoBERT, IndoROBERTa, and XLM-RoBERTa tokenizers were loaded to ensure the input data was segmented and mapped correctly according to each model's vocabulary. The configuration for all models was standardized to match the specific requirements of this study, notably by setting the number of output labels to three (corresponding to the 'negatif', 'netral', and 'positif' sentiment classes). Finally, all computational processes, including model loading and training, were optimized to run on the highest available resource, prioritizing the CUDA (GPU) device, with the CPU as a fallback option.

H. Focal Loss

To effectively address the potential issue of class imbalance and the predominance of easily classified samples within the dataset, the Focal Loss function was implemented as the primary optimization criterion. Focal Loss is designed to down-weight the contribution of easy, well-classified examples during training, compelling the model to focus its learning efforts on hard or misclassified examples [31]. Its structure is defined by the formula:

$$FL(p_t) = -\alpha\,(1- p_t)^{\gamma}.\log(p_t). \qquad (1)$$

In this study, the α (alpha) parameter was used to adjust class weights based on the observed imbalance, specifically set at [2.0, 5.0, 0.3] for the three sentiment classes. Furthermore, the γ (gamma) focusing parameter was fixed at 2.0 to control the rate at which easy examples are down-weighted.

I. Training Strategy and Fine Tune Process

The model training was governed by a strict Training Strategy designed to ensure optimal selection and prevent overfitting. Model performance was continuously monitored with an evaluation step occurring every 50 training steps using the reserved testing set. To preserve valuable progress, a checkpointing process was executed every 100 steps. The crucial Best Model Selection criterion was based on the highest achieved neutral F1-score, as this metric is paramount for distinguishing less extreme opinions accurately in the context of the study. Our save strategy was set to store only the three best performing checkpoints based on this criterion. Furthermore, to prevent resource waste and ensure the final output is the most accurate version, an early stopping mechanism was implemented to load the best model checkpoint at the end of the entire training run.

The core of the optimization involved an iterative Fine-tuning Process spanning 1 to 8 epochs. Within each epoch, the process began with the Forward Pass: the input data was tokenized in batches, the model generated predictions (logits), and the Focal Loss was computed using the predefined class weights. This was immediately followed by the Backward Pass, where gradients were computed and the model weights were updated using the AdamW optimizer step [32]. Throughout the epoch, the model underwent intermittent Evaluation every 50 steps on the test set, computing standard metrics such as accuracy, F1-score, recall, and precision, including detailed per-class metrics. The training state was only saved (checkpointed) when the neutral F1-score showed an improvement, ensuring that only genuinely better models were preserved.

J. Model Evaluation

Upon completion of the fine-tuning process, the final model's effectiveness was rigorously assessed using a Classification Report. This report provided detailed, granular insights into the model's predictive power for each sentiment class by calculating Precision, Recall, and F1-score per class. Furthermore, the evaluation included the calculation of Per-Class Accuracy metrics, specifically breaking down the predictive accuracy for the Negative, Neutral, and Positive categories. This comprehensive evaluation ensures that the model's performance is not only measured overall but is also accurately represented across the potentially imbalanced classes.

The final, best-performing model—selected via the neutral F1-score criterion and the early stopping mechanism—was then permanently stored to ensure its reproducibility and future use. The Save Model process involved preserving all necessary components: the optimized model weights were saved as *pytorch_model.bin* the tokenizer's vocabulary and configuration were stored in *tokenizer.json* and *vocab.txt* the model's architecture parameters were saved in *config.json* and finally, the parameters and settings used during the training process were recorded in *training_args.bin.*

K. Model Test

Following the completion of the fine-tuning process using the designated 70% training data, the resulting optimized model was subjected to a rigorous final test. This essential validation step involved running the model against the unseen 30% testing dataset. The model's generalization capability was then quantitatively assessed by calculating the standard performance metrics: F1-score, Recall, and Precision. This final evaluation confirmed the model's overall effectiveness and its ability to accurately classify sentiment on data it had not previously encountered.

III. RESULT AND DISCUSSION

This section presents the comprehensive performance evaluation of the three fine-tuned pre-trained language models—IndoBERT Base, IndoROBERTa Small, and XLM-RoBERTa Base—based on the metrics obtained from the held-out test dataset as illustrated by figure 5,6, and 7.

A. Fine-Tuning Performance of IndoBERT

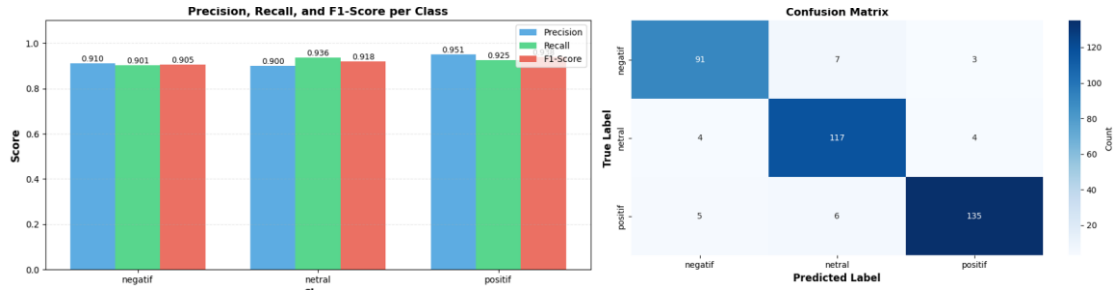

Figure 5. fine tuning results and confussion matrix of IndoBERT model

The IndoBERT model demonstrated excellent performance during the fine-tuning phase, exhibiting high metric scores across all sentiment classes. Specifically, the model achieved the highest F1-score of 0.94 for the Positive class, supported by a precision of 0.95 and a recall of 0.92. The Neutral class also performed strongly, recording an F1-score of 0.92, with a precision of 0.90 and an impressive recall of 0.94. For the Negative class, the model maintained high accuracy, achieving an F1-score of 0.91, derived from a precision of 0.91 and a recall of 0.90. This consistent high performance across all categories indicates that the IndoBERT model successfully learned the distinction between the different sentiment labels on the training dataset.

B. Fine-Tuning Performance of IndoROBERTa Small

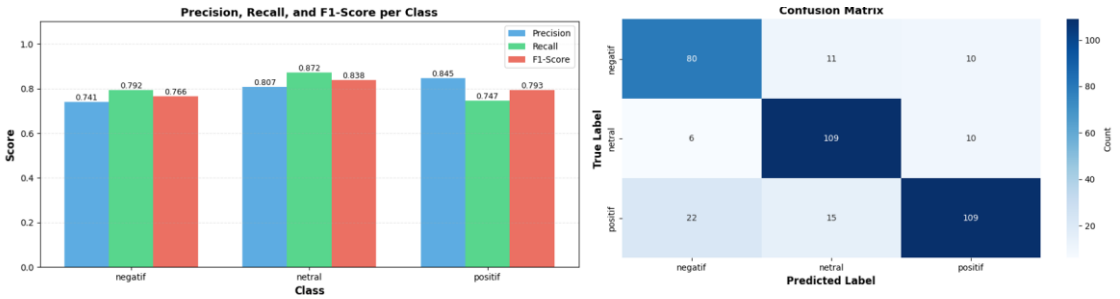

Figure 6. fine tuning results and confussion matrix of IndoROBERTa Small model

The IndoROBERTa Small model demonstrated moderate performance during the fine-tuning phase, showing varied success across the sentiment classes. The highest F1-score was recorded by the Neutral class at 0.84, supported by a precision of 0.81 and a high recall of 0.87. Conversely, the Negative class achieved an F1-score of 0.77, with a precision of 0.74 and a recall of 0.79. The Positive class recorded an F1-score of 0.79, primarily limited by a lower recall of 0.75 despite having the highest precision among the classes at 0.84. Overall, the performance suggests the model faced slightly more difficulty in classifying extreme sentiments compared to the neutral category on the training data.

C. Fine-Tuning Performance of XLM-RoBERTa

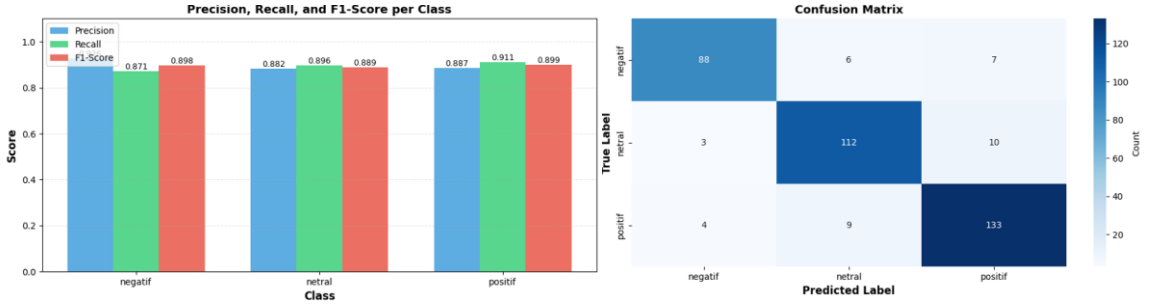

Figure 7. fine tuning results and confussion matrix of XLM-RoBERTa model

The XLM-RoBERTa Base model showed strong and balanced performance during the fine-tuning stage. The Negative and Positive classes achieved the highest F1-scores of 0.90, demonstrating the model's high capability to distinguish extreme opinions. Specifically, the Negative class was supported by a high precision of 0.93 and a recall of 0.87. The Positive class maintained good performance with a precision of 0.89 and a recall of 0.91. The Neutral class was also successfully classified, achieving an F1-score of 0.89, based on a precision of 0.88 and a recall of 0.90. This demonstrates that the multilingual model effectively learned the sentiment patterns within the Indonesian review data. The performance evaluation of the three models—IndoBERT, IndoROBERTa Small, and XLM-RoBERTa—on the test dataset is presented in detail.

A. Test Performance of IndoBERT Base

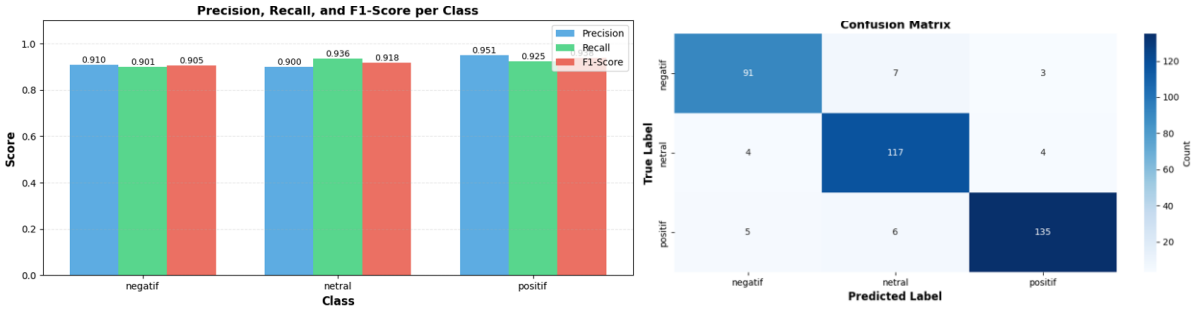

Figure 8. unseen test dataset results and Confusion matrix using the IndoBERT model

On the unseen test dataset, the IndoBERT Base model maintained a strong and reliable performance, confirming its generalization capabilities. Illustrated by figure 8, the highest F1-score of 0.90 was achieved by the Negative class, driven by an excellent precision of 0.94 and a recall of 0.86. The Positive class also performed highly, registering an F1-score of 0.89, with a precision of 0.89 and a recall of 0.90. The Neutral class recorded an F1-score of 0.88, supported by a precision of 0.86 and a recall of 0.90. The overall high and balanced F1-scores across all three sentiment classes validate the model's robustness and suitability for classifying new, real-world *pesantren* reviews.

B. Test Performance of IndoROBERTa Small

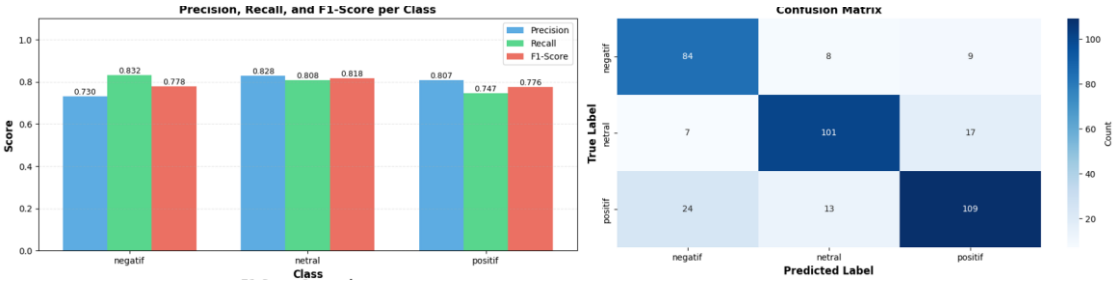

Figure 9. unseen test dataset results and Confusion matrix using the IndoROBERTa Small model

When tested on the unseen dataset, the IndoROBERTa Small model showed moderate generalization capabilities as illustrated by figure 9, with metrics lower than the other models evaluated. The highest F1-score of 0.82 was achieved by the Neutral class, supported by a precision of 0.83 and a recall of 0.81. Conversely, the Negative and Positive classes recorded the lowest F1-scores, both at 0.78. Specifically, the Negative class had a precision of 0.73 and a high recall of 0.83, while the Positive class achieved an F1-score of 0.78 with a precision of 0.81 and a lower recall of 0.75. This performance indicates that the IndoROBERTa Small model, despite its efficiency, struggled to maintain high accuracy when classifying extreme sentiments on new data.

C. Test Performance of XLM-RoBERTa Base

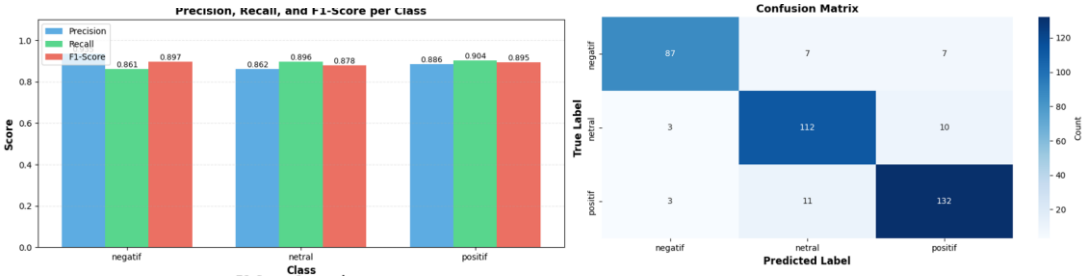

Figure 10. unseen test dataset results using the XLM-RoBERTa Base model

When evaluated on the unseen test dataset, the XLM-RoBERTa Base model demonstrated strong generalization performance across all classes. Illustrated by figure 10, the highest F1-score of 0.90 was achieved by the Negative class, driven by an excellent precision of 0.94 and a recall of 0.86. The Positive class also performed very well, recording an F1-score of 0.89, with a precision of 0.89 and a recall of 0.90. The Neutral class was successfully classified with an F1-score of 0.88, supported by a precision of 0.86 and a high recall of 0.90. These consistent results confirm the model's ability to maintain high performance when applied to new, real-world data outside of the training environment.

D. Discussion

The experimental results confirm that IndoBERT Base is the superior model for this dataset, achieving an accuracy of 92%. A critical comparison suggests that monolingual models pre-trained specifically on Indonesian corpora (IndoBERT) possess a deeper contextual grasp of local school-related nuances compared to the broader, yet less precise, multilingual generalization of XLM-RoBERTa.

The implementation of Focal Loss played a decisive role in mitigating the class imbalance noted in the dataset (4,000 positive vs. 1,165 negative reviews). By utilizing the focusing parameter $\Upsilon = 2.0$ and specific class weights $\alpha = [2.0, 5.0, 0.3]$, the models were forced to learn from "hard examples" (negative and neutral sentiments) rather than being overwhelmed by the majority positive class. This is evidenced by the balanced F1-scores across all categories, preventing the common "majority bias" found in standard cross-entropy training.

Theoretically, this research validates the scalability of IndoBERT variants for specialized Indonesian sub-domains involving non-standard language. Practically, pesantren administrators can utilize this model as an automated auditing tool to monitor institutional reputation on Google Maps. Negative sentiments identified (e.g., keywords like "kotor", "mahal", or "fasilitas kurang" in Figure 3) provide direct, actionable feedback for facility and service improvements.

Despite the high accuracy, this study is limited by its geographic focus on the island of Java. Linguistic variations and dialects from other regions in Indonesia may influence the model's performance on a national scale. Furthermore, the manual labeling process, while accurate, remains a bottleneck for scaling the dataset significantly in future iterations.

IV. CONCLUSIONS

Based on the quantitative results, the IndoBERT Base model is conclusively identified as the optimal architecture for the sentiment analysis of Google Maps reviews concerning pesantren in Java. The model achieved a robust and consistent accuracy of 92% on both the fine-tuning and testing datasets, which strongly indicates excellent generalization capabilities and the absence of significant overfitting. This high performance validates the chosen methodology, particularly the integration of comprehensive text preprocessing and the strategic implementation of the Focal Loss function to maintain balanced F1-scores across all sentiment classes—Negative (0.91), Neutral (0.92), and Positive (0.94). The primary contribution of this research is the establishment of a specialized sentiment classification framework for the religious-educational domain, which effectively addresses class imbalance and linguistic nuances. Unlike traditional methods, this study provides a robust tool for pesantren stakeholders to monitor institutional reputation and gain actionable insights from public feedback in a sensitive educational context. For future research, the focus should shift toward building a larger, more diverse, and meticulously curated dataset. Improving the volume and geographic diversity of the data will be key to achieving state-of-the-art results and ensuring broader generalization across various review types in Indonesia.

REFERENCES

[1] H. Nashihin, N. Aziz, I. Z. Adibah, N. Triana, and Q. Robbaniyah, "KONSTRUKSI PENDIDIKAN PESANTREN BERBASIS TASAWUF-ECOSPIRITUALISM DAN ISU LINGKUNGAN HIDUP," 2022, doi: 10.30868/ei.v11i01.2794.

[2] A. Hasan and A. Asyari, "Tantangan Sistem Pendidikan Pesantren di Era Modern," 2022. [Online]. Available: https://ejournal.iaisyarifuddin.ac.id/index.php/risalatuna

[3] Mita Silfiyasari and Ashif Az Zhafi, "Peran Pesantren dalam Pendidikan Karakter di Era Globalisasi," *Jurnal Pendidikan Islam Indonesia*, vol. 5, no. 1, pp. 127–135, Oct. 2020, doi: 10.35316/jpii.v5i1.218.

[4] M. R. Kamal, K. Noviyanto, H. Subhan, and L. Afiana, "Sentiment Analysis on Social Media and Stakeholders about Negative Issues Among Islamic Boarding School Community In Indonesia," *Jurnal Theologia*, vol. 35, no. 2, pp. 261–280, Dec. 2024, doi: 10.21580/teo.2024.35.2.23164.

[5] Y. A. Laghbi and M. Al Dhoayan, "Examining how customers perceive community pharmacies based on Google maps reviews: Multivariable and sentiment analysis," *Exploratory Research in Clinical and Social Pharmacy*, vol. 15, p. 100498, Sep. 2024, doi: 10.1016/j.rcsop.2024.100498.

[6] P. Phuangsuwan, S. Siripipatthanakul, P. Limna, and N. Pariwongkhuntorn, "The impact of Google Maps application on the digital economy," *Phuangsuwan, P., Siripipatthanakul, S., Limna, P., & Pariwongkhuntorn*, no. 2024, pp. 192–203, 2024.

[7] B. Liu, *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[8] M. Gharzouli, A. K. Hamama, and Z. Khattabi, "Topic-based sentiment analysis of hotel reviews," *Current Issues in Tourism*, vol. 25, no. 9, pp. 1368–1375, May 2022, doi: 10.1080/13683500.2021.1940107.

[9] E. Damayanti, A. V. Vitianingsih, S. Kacung, H. Suhartoyo, and A. Lidya Maukar, "Sentiment Analysis of Alfagift Application User Reviews Using Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) Methods," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 4, no. 2, pp. 509–521, Jun. 2024, doi: 10.51454/decode.v4i2.478.

[10] M. Atif, "An Enhanced Framework for Sentiment Analysis of Students' Surveys: Arab Open University Business Program Courses Case Study," *Bus. Econ. J.*, vol. 09, no. 01, 2018, doi: 10.4172/2151-6219.1000337.

[11] M. Ilona Junide Bria, A. Vega Vitianingsih, A. Lidya Maukar, S. Yuliani, and A. Info, "Sentiment Analysis of User Reviews on Maxim Application Using the Long Short-Term Memory (LSTM) Methods," vol. 5, pp. 941–952, 2025, doi: 10.51454/decode.v5i3.1257.

[12] F. M. Apriansyah, T. I. Ramadhan, C. R. Hidayat, and A. K. Wijaya, "Perbandingan IndoBERT dan IndoROBERTa Untuk Analisis Sentimen Pada Film Dokumenter Dirty Vote," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 3, pp. 593–605, Jul. 2025, doi: 10.30591/jpit.v10i3.8607.

[13] N. Rihaadatul'Aisy and E. W. Pamungkas, "Sentiment Analysis of Indonesian News Texts Using IndoBERT and IndoROBERTa ," in *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/SIML65326.2025.11080939.

[14] M. Lalthangmawii and T. D. Singh, "Sentiment analysis of Mizo using lexical features in low resource based models," *Natural Language Processing Journal*, vol. 13, p. 100181, Dec. 2025, doi: 10.1016/j.nlp.2025.100181.

[15] M. Ridha, D. Nurjanah, and M. Rakha, "Multilabel Classification Abusive Language and Hate Speech on Indonesian Twitter using Transformer Model: IndoBERT weet & IndoROBERTa ," in *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 2024, pp. 48–54.

[16] F. Dinarta and A. Wicaksana, "Enhanced Hate Speech Detection in Indonesian-English Code-Mixed Texts Using XLM-RoBERTa," *Informatica*, vol. 49, no. 21, 2025.

[17] F. Basbeth, "Klasifikasi Emosi Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan DistilBERT," Universitas Islam Indonesia, 2024.

[18] H. Cuesta and S. Kumar, *Practical data analysis*. Packt Publishing Ltd, 2016.

[19] W. Van Atteveldt, M. A. C. G. der Velden, and M. Boukes, "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Commun. Methods Meas.*, vol. 15, no. 2, pp. 121–140, 2021.

[20] J. Zhu, X. Zhao, Y. Sun, S. Song, and X. Yuan, "Relational data cleaning meets artificial intelligence: A survey.," *Data Sci. Eng.*, vol. 10, no. 2, 2025.

[21] N. W. A. S. Aprilia and A. R. Isnain, "Analisis Sentimen Terhadap Media Sosial Twitter dengan Kasus Kampanye Anti-Korupsi di Indonesia Menggunakan Naive Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, p. 695, Apr. 2024, doi: 10.30865/mib.v8i2.7582.

[22] W. Bourequat and H. Mourad, "Sentiment analysis approach for analyzing iPhone release using support vector machine," *International Journal of Advances in Data and Information Systems*, vol. 2, no. 1, pp. 36–44, 2021.

[23] H. Mohamed Zakir and S. Vinila Jinny, "A Comparative Study on Data Cleaning Approaches in Sentiment Analysis," *Advances in Communication Systems and Networks: Select Proceedings of ComNet 2019*, pp. 421–431, 2020.

[24] C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, and M. Cannataro, "Sentiment analysis for mining texts and social networks data: Methods and tools," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 1, p. e1333, 2020.

[25] E. Prasetio Widhi, P. Sholihin, A. Musthafa, and N. Marantika, "Journal of Artificial Intelligence and Engineering Applications Sentiment Analysis of Public Trust Towards Islamic Boarding School on Social Media Using Machine Learning Method," 2025. [Online]. Available: https://ponpeskeren.id,

[26] A. Erkan and T. Güngör, "Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification," *IEEE Access*, vol. 11, pp. 134951–134968, 2023.

[27] A. W. Pradana and M. Hayaty, "The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 375–380, 2019.

[28] H. A. Shehu *et al.*, "Deep sentiment analysis: a case study on stemmed Turkish twitter data," *Ieee Access*, vol. 9, pp. 56836–56854, 2021.

[29] F. Farias, T. Ludermir, and C. Bastos-Filho, "Similarity Based Stratified Splitting: an approach to train better classifiers," *arXiv preprint arXiv:2010.06099*, 2020.

[30] N. Varshney and S. Singh, "Enhancing Diabetes Prediction: A comparative analysis of Train-Test Split and Stratified 10-Fold Cross-Validation with SMOTE Integration," in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 2024, pp. 1345–1351.

[31] P. Basu, S. Tiwari, J. Mohanty, and S. Karmakar, "Multimodal sentiment analysis of# metoo tweets using focal loss (grand challenge)," in *2020 IEEE sixth international conference on multimedia big data (BigMM)*, 2020, pp. 461–465.

[32] R. Llugsi, S. El Yacoubi, A. Fontaine, and P. Lupera, "Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito," in *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, 2021, pp. 1–6.