Jurnal TRANSFORMATIKA
Vol.23, No.1, Juli 2025, pp. 85 - 95
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews

Dhendra Marutho[1*], Victor Gayuh Utomo[2]
[1]Informatika, Universitas Muhammadiyah Semarang
Jl. Kedung mundu Raya Semarang 50273, 024-76740296, e-mail: dhendra@unimus.ac.id
[2]Sistem Informasi, Universitas Semarang
Soekarno-Hatta Semarang 50196, 024-6702757, e-mail: victor@usm.ac.id

ARTICLE INFO

History of the article :
Received 25 Mei 2025
Received in revised form 8 Juli 2025
Accepted 8 Juli 2025
Available online 16 Juli 2025

Keywords:

Sentiment Analysis; Natural Language Processing; IndoBERT; Transformers Model; Data Mining

*** Correspondece:**

Telepon:
+628562655511

E-mail:
dhendra@unimus.ac.id

ABSTRACT

This study aims to benchmark the performance of five sentiment classification models for Indonesian e-government service reviews. We used a hybrid labeling approach combining rating-based heuristics with manual validation to improve label accuracy. The five models were selected to represent both transformer-based architectures (IndoBERT, mBERT, XLM-R) and classical deep learning baselines (CNN with Word2Vec, BiLSTM with GloVe) for comprehensive comparison. The sentiment categories classified include positive, negative, and neutral classes. Experimental results show that IndoBERT outperformed other models, achieving an accuracy of 88.1% and an F1-score of 0.881, indicating the potential of language-specific transformer models for automated public feedback analysis.

1. INTRODUCTION

The rapid advancement of digital governance has led to increased reliance on e-government services in many developing countries, including Indonesia. These services range from online taxation and licensing systems to digital identity management platforms. They play a crucial role in facilitating public administration and enhancing citizen engagement. However, the success of such digital initiatives depends not only on their technological implementation but also on their perceived quality, usability, and trustworthiness by end-users. Measuring public sentiment toward these services can thus offer valuable insights for both system improvements and policy making.

Sentiment analysis, also known as opinion mining, has emerged as a powerful tool for interpreting user-generated content and assessing public perception. While earlier studies have relied heavily on traditional machine learning techniques such as Support Vector Machines and Naïve Bayes classifiers. The emergence of deep learning and transformer-based models—

particularly Bidirectional Encoder Representations from Transformers (BERT) [1] has significantly improved sentiment classification performance across many languages and domains [2], [3].

In the context of Indonesian language processing, IndoBERT—an adaptation of BERT pre-trained on over 4 billion Indonesian tokens—has demonstrated strong performance in a range of tasks including sentiment analysis, emotion classification, and aspect-based opinion mining [4], [5]. Despite these achievements, existing research often focuses on sectors such as healthcare applications [3], product reviews [6], or political opinions [7], with relatively few studies exploring its application in evaluating public sentiment toward government digital services.

While prior sentiment analysis research in Indonesia has largely focused on domains such as e-commerce, transportation (e.g., LRT Jabodebek) [8] little to no attention has been paid to public-facing digital services managed by the national police, such as SIM online renewal, vehicle tax (STNK) management, and Samsat e-payment platforms. These services are critical for citizen mobility, legal compliance, and daily governance, yet they remain underexamined from a digital sentiment perspective.

This study addresses this gap by introducing a curated dataset of user reviews from NEWSAKPOLE, a digital public service application developed by the Regional Revenue Agency (Bapenda) of Central Java for vehicle tax payment and licensing. We benchmark the performance of several transformer-based models, including IndoBERT, mBERT, and XLM-R, as well as classical deep learning baselines such as CNN with Word2Vec and BiLSTM with GloVe, on this domain-specific dataset. In doing so, we aim to (1) evaluate the effectiveness of modern NLP techniques for classifying Indonesian sentiment in public service reviews and (2) provide actionable insights for improving citizen trust and digital service delivery in Indonesia's e-government ecosystem.

2. RELATED WORK

2.1. SENTIMENT ANALYSIS IN PUBLIC SERVICES AND E-GOVERNMENT

Sentiment analysis has become a prominent method for understanding public opinion in sectors such as transportation, healthcare, education, and governance. In Indonesia, while research in sentiment analysis has been widely conducted in commercial and health domains, studies focusing on public services remain limited. One notable exception is the work on LRT Jabodebek, where IndoBERT was utilized alongside lexicon-based and BERTopic methods to analyze sentiments from Twitter, Instagram, and YouTube comments [8]. The study reported an F1-score of 84.13% and showed that the majority sentiment was negative (55.9%), highlighting a dissatisfaction trend. However, such studies are domain-specific and fail to generalize across the broader spectrum of digital public services such as taxation, civil registration, and online complaint platforms.

Another relevant application of sentiment analysis is in the healthcare sector. Yulianti [3] conducted a study on Indonesian health applications such as Halodoc and Alodokter using IndoBERT, achieving high accuracy up to 96%. Although impactful, these studies focus on specific service types and do not benchmark across models nor target services operated by state institutions like the national police (POLRI), which offer services including SIM renewal and STNK e-payment.

2.2. INDOBERT AND TRANSFORMER-BASED MODELS IN INDONESIAN NLP

IndoBERT is a transformer-based language model pre-trained specifically for Bahasa Indonesia on a 4B-token corpus. It has been successfully fine-tuned for multiple downstream NLP tasks including sentiment classification, emotion detection, paraphrase identification, and summarization [4]. Studies consistently show that IndoBERT outperforms traditional models such as Naïve Bayes, SVM, and even CNN-based models, particularly when applied to informal or short-text data [2], [9]. For instance, a study comparing IndoBERT to CNN, RCNN, and BiGRU

architectures for analyzing student sentiment found that IndoBERT achieved the highest accuracy and generalization capability [6].

Recent research also explores hybrid architectures that combine IndoBERT embeddings with CNN or GRU layers to enhance classification performance in noisy or informal datasets [5]. These approaches demonstrate that context-aware embeddings, such as those produced by transformer models, significantly outperform older vector space models like Word2Vec or TF-IDF. In addition to IndoBERT, other transformer models such as mBERT and XLM-R have been applied for multilingual sentiment classification. However, their performance on Bahasa Indonesia datasets is often suboptimal due to the lack of language-specific pretraining [10]. Meanwhile, classical deep learning architectures such as CNN with Word2Vec and BiLSTM with GloVe continue to be used as benchmarking baselines because of their simplicity and interpretability. Nonetheless, these models generally underperform compared to transformer-based models, especially when applied to informal or domain-specific datasets [2], [5]

2.3. BENCHMARKING IN LOW-RESOURCE NLP

Despite the growing use of IndoBERT, benchmarking efforts in low-resource languages, including Bahasa Indonesia, remain limited. Most comparative studies focus on high-resource languages or utilize generic datasets such as product or movie reviews, which do not reflect the complexity of public service discourse. Bania et al. [11], for example, applied TF-IDF and traditional classifiers like SVM and Random Forest to COVID-19-related tweets, which, while relevant, lack the semantic depth captured by contextual models.

Furthermore, although classical models continue to be used due to their simplicity and interpretability, their effectiveness in handling ambiguous, sarcastic, or implicit sentiment remains low [2]. With the increasing availability of pre-trained transformer models and access to GPU computing, the urgency of conducting domain-specific benchmarks in underrepresented languages becomes even more apparent.

To the best of our knowledge, no prior study has attempted to benchmark transformer-based models—including IndoBERT—on Indonesian digital public service reviews related to POLRI platforms. This research addresses that gap by introducing a new dataset and performing comparative evaluations that reflect real-world government service scenarios.

3. METHODOLOGY

This study aims to benchmark the performance of IndoBERT and other transformer-based models in classifying sentiment from user reviews of an Indonesian digital public service application. The research workflow includes five main stages: data acquisition, preprocessing, sentiment labeling, model implementation, and evaluation.

3.1. DATA COLLECTION AND PREPROCESSING

The dataset was collected by scraping user reviews of the NEWSAKPOLE application from the Google Play Store. This application, managed by the Regional Revenue Agency (Bapenda) of Central Java, facilitates online services such as vehicle tax payments and license renewals.

To automate data acquisition, we used the *google-play-scraper* Python library [12]. The script was configured to extract up to 10,000 user reviews, capturing the following fields: review content, rating (1–5 stars), timestamp, thumbs up count, and review version. The query was targeted specifically for Indonesian-language reviews**.**

DOI : https://doi.org/ 10.26623/transformatika.v23i1.12095

	review_text	category
0	erorr	negative
1	Simple dan bermanfaat	positive
2	Pernah dulu bayar lewat aplikasi sakpole yg pertama,itu memudahkan sekali saya bayar pajak kendaraan ketika di perantauan. Sekarang di NewSakpole kenapa jadi rada ribet ya dengan foto macam2,terus selalu muncul notif belum waktunya bayar karena waktunya masih lebih dari 60 hari padahal waktunya tinggal 50 hari ke jatuh tempo. Cek pembaruan aplikasi gak ada pembaruan. Saya mau bayar cepat mumpung ada duit padahal. Org bijak taat pajak,udah Bijak bgt kan gue..	neutral
3	Buat ngecek pajek susah sekali	negative
4	Mbuh lah AngeL..	negative

Figure 1. Example of User Reviews from NEWSAKPOLE Dataset with Sentiment Labels

Each review was preprocessed and labeled as positive, negative, or neutral using a rating-based heuristic and manual validation.

3.2. TEXT PREPROCESSING

Raw reviews were cleaned and normalized using a standard NLP pipeline designed for informal Indonesian text. The preprocessing steps included lowercasing all characters [13], normalizing slang expressions (for example, converting "ga" to "tidak"), tokenizing the text using the IndoNLP tokenizer [14], removing stopwords using an Indonesian stopword list [15], and applying stemming with the Sastrawi stemmer [16]. Formally, let a raw review be defined as:
Formally, let a raw review be defined as:

$$r = \{w_1, w_2, \ldots, w_n\} \quad (1)$$

After preprocessing, the cleaned version becomes:

$$T` = \{t_1, t_2, \ldots, t_n\}, with\ k \leq \text{n} \quad (2)$$

3.3. SENTIMENT LABELING

A semi-supervised labeling approach was adopted in this study. Reviews with ratings of 1–2 were labeled as Negative (label = 0), while reviews with ratings of 4–5 were labeled as Positive (label = 1). Reviews with a rating of 3 were excluded to reduce ambiguity in sentiment interpretation. To ensure label reliability, 20% of the reviews were manually validated by human annotators to check alignment between the assigned rating and the actual textual sentiment.

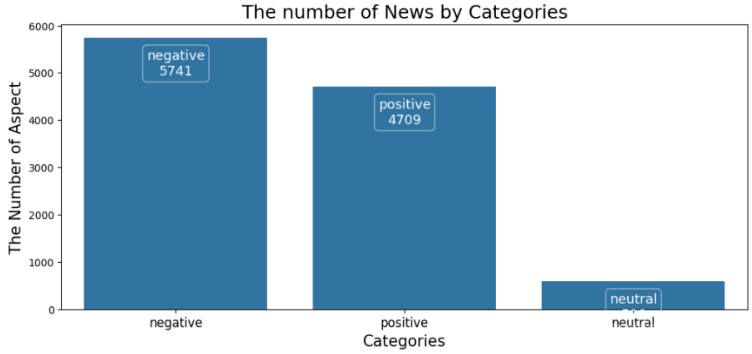

Figure 2. Distribution of Sentiment Categories in the NEWSAKPOLE Dataset

The dataset is dominated by negative reviews (5,741), followed by positive (4,709) and a smaller proportion of neutral reviews (approximately 600). This imbalance presents challenges in model training, particularly for the neutral class.

3.4. MODEL ARCHITECTURE AND TRAINING

We benchmarked the following models:

Table 1. Description of Sentiment Classification Models Used in This Study

Model Type	Description
IndoBERT	Pretrained BERT model for Indonesian
mBERT	Multilingual BERT
XLM-RoBERTa	Cross-lingual model trained on 100+ languages
CNN + Word2Vec	Baseline convolutional model
BiLSTM + GloVe	Recurrent model baseline

The classification layer for BERT-based models uses the hidden state of the [CLS] token, followed by a softmax activation:

$$\hat{y} = softmax(W.H_{[CLS]} + b) \tag{3}$$

All models were trained using the **Cross-Entropy Loss**:

$$L_{ce} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}.Log(\hat{y}_{i,c}) \tag{4}$$

Training parameters were set to ensure effective learning and fair comparison across models. The optimizer used was AdamW with a learning rate of 2×10^{-5}. Each model was trained with a batch size of 16 for 10 epochs. The dataset was split into 80% for training, 10% for validation, and 10% for testing to evaluate model performance consistently.

3.5. EVALUATION METRICS

Model performance was evaluated using standard classification metrics:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

F1:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

F1 Score:

DOI : https://doi.org/ 10.26623/transformatika.v23i1.12095

$$F1_{\text{score}} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{9}$$

Where *TP, TN, FP, FN* are the confusion matrix elements and C=3

3.6. SUMMARY WORKFLOW

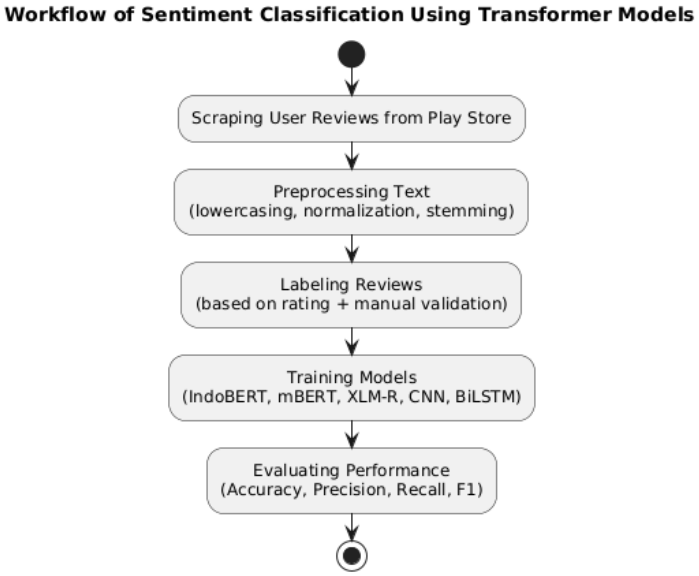

Figure 3. Workflow of Sentiment Classification using Transformers Models

Figure 3 illustrates the workflow of the sentiment classification process applied in this study. The pipeline begins with the collection of user reviews from the Google Play Store using automated scraping methods. The extracted raw data, written in informal Indonesian, undergoes a preprocessing stage that includes lowercasing, slang normalization, tokenization, stopword removal, and stemming to ensure consistent linguistic structure.

In the subsequent stage, the reviews are labeled using a hybrid approach. Reviews are automatically labeled based on star ratings (1–2 as negative, 4–5 as positive), with a subset of data manually validated to improve labeling accuracy and reduce misclassification noise.

The labeled dataset is then used to train several models, including transformer-based architectures such as IndoBERT, mBERT, and XLM-R, as well as classical deep learning baselines like Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. Each model is trained on the same data split to ensure fair comparison. Finally, the trained models are evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. This comprehensive workflow enables the benchmarking of multiple approaches to identify the most effective model for sentiment classification in Indonesian e-government service reviews.

4. RESULTS AND DISCUSSION

4.1. MODEL PERFORMANCE COMPARISON

This study evaluates five models for sentiment classification on Indonesian-language user reviews from the NEWSAKPOLE application. These models include three transformer-based approaches—IndoBERT, XLM-RoBERTa (XLM-R), and mBERT—as well as two classical

deep learning baselines: Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM).
Each model was trained and validated using the same dataset, with 80% of the data allocated for training, 10% for validation, and 10% for testing. Evaluation was based on accuracy, precision, recall, and F1-score.

Table 2. Performance comparison of classification models

Model	Accuracy	Precission	Recall	F1-Score
IndoBERT	**0.881**	0.8813	0.8813	0.8813
XLM-Roberta	0.880	0.880	0.880	0.880
mBERT	0.868	0.868	0.868	0.868
CNN+Word2Vec	0.845	0.828	0.845	0.836
BiLSTM+Glove	0.841	0.827	0.841	0.833

presents the comparative results. IndoBERT achieved the highest F1-score of 0.881 and accuracy of 88.1%, closely followed by XLM-R (F1 = 0.880) and mBERT (F1 = 0.868). Classical models such as CNN and BiLSTM underperformed relative to transformer models, yielding F1-scores of 0.845 and 0.841, respectively. These results suggest that adopting transformer-based models like IndoBERT can significantly improve automated analysis of citizen feedback in public service applications. Such integration can support faster policy responses, enhance service quality, and strengthen public trust in e-government systems.

4.2 BEST PERFORMING MODEL: INDOBERT

IndoBERT consistently outperformed other models across all metrics. Its success can be attributed to pretraining on over 4 billion Indonesian tokens, enabling the model to better understand the linguistic nuances of Bahasa Indonesia, particularly in informal review contexts. Compared to multilingual models like XLM-R and mBERT, IndoBERT showed higher recall and precision, demonstrating its ability to detect both positive and negative sentiment reliably. Classical models, though effective in many text classification tasks, lacked the contextual depth to manage noisy, colloquial user reviews as effectively.
The superiority of IndoBERT suggests that using language-specific pretrained models yields better performance in low-resource language settings.

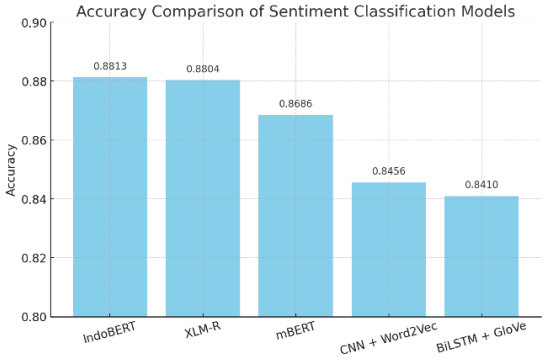

Figure 4. Accuracy Comparion of Sentiment Classfication Models

4.3 CONFUSION MATRIX AND ERROR ANALYSIS

To gain deeper insights into the classification behavior of each model, confusion matrices were analyzed for IndoBERT, XLM-R, mBERT, CNN, and BiLSTM, as shown in Figures

DOI : https://doi.org/ 10.26623/transformatika.v23i1.12095

5(a)–5(e). The analysis aims to reveal patterns of misclassification and performance disparities across sentiment classes.

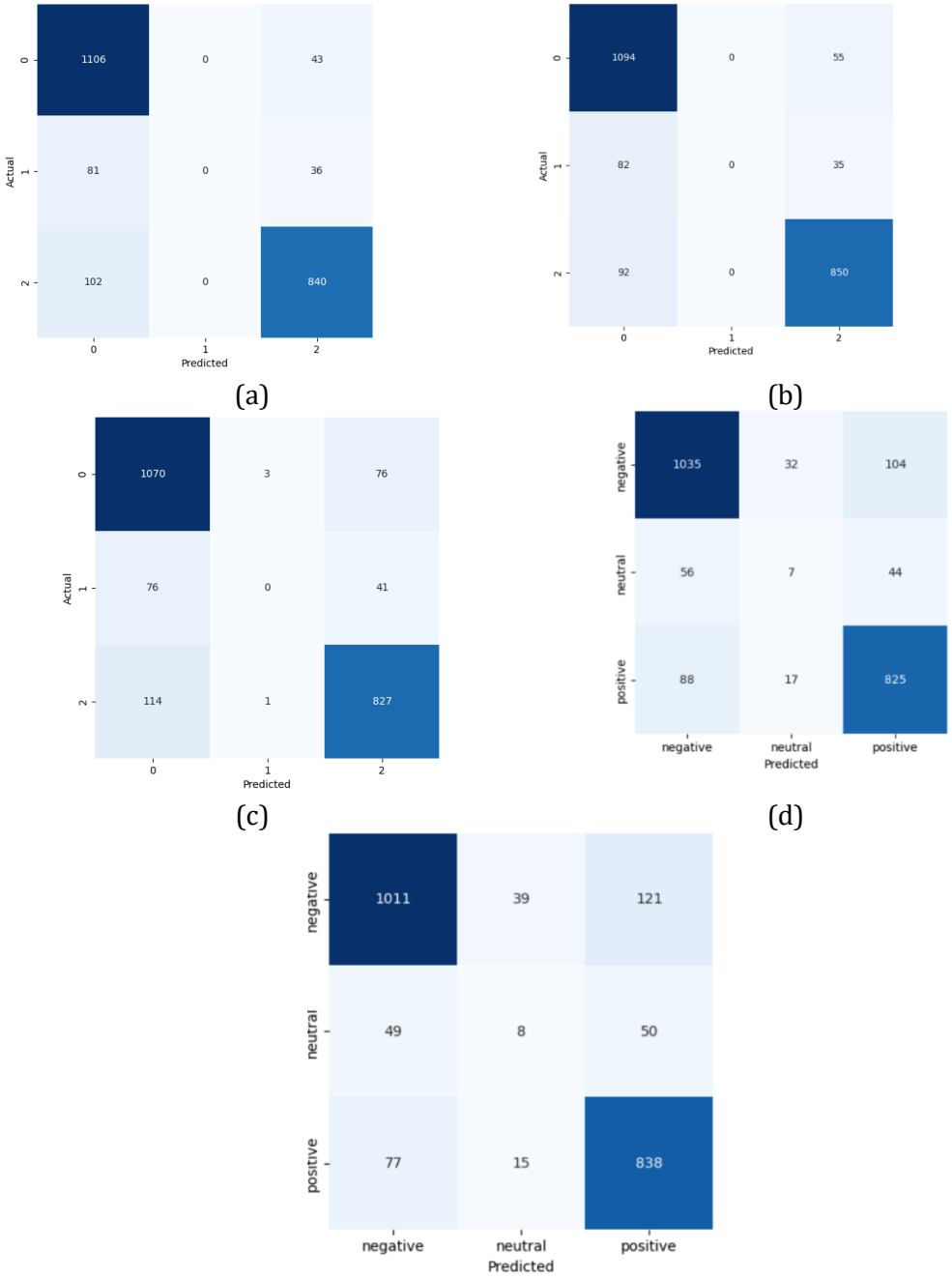

Figure 5 Confusion matrices of sentiment classification models. (a) IndoBERT, (b) XLM-RoBERTa, (c) mBERT, (d) CNN + Word2Vec, (e) BiLSTM + GloVe.

The confusion matrix for IndoBERT Figure 5(a) reveals strong performance in both the negative and positive classes, with only minor misclassifications. Most notably, IndoBERT misclassified 102 positive reviews as negative, likely due to ambiguous expressions, sarcasm, or low context cues. The neutral class exhibited relatively higher misclassification, reflecting a broader challenge in distinguishing truly neutral sentiment within informal Indonesian reviews. Similarly, XLM-R Figure 5(b) demonstrates comparable patterns, with slightly higher confusion between classes compared to IndoBERT. Although XLM-R maintains a balanced precision and recall, its multilingual nature and token generalization possibly reduce its capacity to resolve domain-specific sentiment cues effectively.

For mBERT Figure 5(c), the matrix indicates better balance than CNN or BiLSTM yet still underperforms compared to IndoBERT. There is notable confusion between positive and negative classes, though less severe than traditional deep learning baselines. In contrast, CNN and BiLSTM exhibit more severe misclassification trends, particularly in the neutral class. As seen in Figure 4(d), CNN correctly identified only 7 of 107 neutral reviews, while BiLSTM correctly classified only 8 Figure 5(e). Furthermore, both models frequently misclassified positive sentiment as negative 104 cases in CNN and 121 in BiLSTM indicating a lack of semantic discrimination when contextual polarity is subtle.

These results suggest that traditional models, relying on static word embeddings (Word2Vec, GloVe), struggle to capture complex semantic relationships and nuances in sentiment. Their linear or sequential architectures are ill-equipped to deal with implicit sentiment or noisy, unstructured text common in public service feedback.

4.4 DISCUSSION AND INTERPRETATION

The comparative evaluation clearly shows the superiority of transformer-based models for sentiment classification tasks in Bahasa Indonesia, especially within public service domains. IndoBERT delivered the best results across all metrics, supported by its pretraining on a massive Indonesian corpus. Its strong contextual embedding capability enabled accurate classification of informal, noisy, and short-form user reviews. XLM-R, although multilingual, followed closely behind, indicating its generalizability despite lacking domain-specific tuning. mBERT presented consistent performance but remained inferior to IndoBERT and XLM-R. This aligns with previous research noting that mBERT, trained on multilingual corpora without focus on Indonesian, lacks fine-grained semantic precision in local contexts. On the other hand, CNN and BiLSTM showed lower accuracy and F1-scores, and their confusion matrices highlighted critical weaknesses. These models frequently misclassified neutral and positive sentiment as negative, which may stem from their inability to model negation, irony, or mixed tones. The neutral class emerged as the most difficult to classify across all models. This difficulty is likely caused by the absence of strong emotional indicators in truly neutral texts, the overlap of lexical features with both positive and negative sentiments, and the possibility of sarcastic or low-engagement responses from reviewers. To improve classification in future research, incorporating attention mechanisms or hierarchical models could help better capture contextual cues. Additionally, exploring contextual sentiment lexicons specific to public services may enhance model understanding of domain-related expressions. Manual annotation refinement for neutral samples is also recommended to improve labeling quality and reduce misclassification. In conclusion, the transformer models—particularly IndoBERT—demonstrate high potential for practical deployment in Indonesian e-government applications. Their robust performance offers valuable insights for improving public feedback systems, guiding system enhancements, and strengthening trust in digital governance services.

DOI : https://doi.org/ 10.26623/transformatika.v23i1.12095

5. CONCLUSION AND FUTURE WORK

This study investigated the effectiveness of transformer-based and deep learning models for sentiment classification on Indonesian-language user reviews of public service applications, using a curated dataset from the NEWSAKPOLE platform. The objective was to benchmark multiple models—IndoBERT, XLM-R, mBERT, CNN, and BiLSTM—on their ability to classify user sentiment in the context of e-government services, with a specific focus on low-resource language settings and informal public feedback. The experimental results show that IndoBERT outperformed all other models, achieving an F1-score of 0.881, followed closely by XLM-R (0.881) and mBERT (0.869). These findings affirm that language-specific transformer models are better suited for sentiment classification tasks in Bahasa Indonesia, especially when dealing with noisy and informal user-generated content. In contrast, classical deep learning models such as CNN and BiLSTM yielded significantly lower performance (F1-scores of 0.836 and 0.833, respectively), and struggled particularly with identifying neutral sentiment. Confusion matrix analysis further confirmed that transformer-based models are more effective at handling ambiguous or mixed sentiment, whereas CNN and BiLSTM exhibited high misclassification rates, especially for neutral reviews. These patterns highlight the importance of context-aware modeling and the limitations of static embedding-based architectures in nuanced sentiment analysis.

IMPLICATIONS

The results of this study are directly applicable to improving public feedback analysis in Indonesian e-government systems. By integrating IndoBERT or similar transformer models, stakeholders can automate the monitoring of citizen satisfaction and complaints, enabling faster policy responses and enhancing public trust.

LIMITATIONS

This study is limited to a single public service domain (vehicle tax and licensing), and the sentiment labels were derived through rating heuristics and manual validation, which may introduce labeling bias. Additionally, the neutral class remained underrepresented and harder to predict, suggesting a need for more balanced and diverse data.

FUTURE WORK

Several directions can be explored in future research. Expanding the dataset to include multiple government service applications and broader regional contexts would increase generalizability. Investigating aspect-based sentiment analysis (ABSA) can help identify specific service components that influence public opinion. Applying prompt-based learning or parameter-efficient fine-tuning (PEFT) techniques may further enhance transformer model performance even with minimal data. Additionally, exploring multi-modal sentiment classification using text, audio, and image reviews (such as screenshots of service complaints) can improve the robustness and applicability of these models. By addressing these challenges, sentiment analysis approaches can become more reliable, actionable, and better support responsive and citizen-centric governance.

REFERENCES

[1] D. Marutho, Muljono, S. Rustad, and Purwanto, "Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 1, p. 100211, 2024, doi: 10.1016/j.joitmc.2024.100211.

[2] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, Mar. 2025, doi: 10.1007/s13278-025-01439-6.

[3] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, 2023, doi: 10.14569/ijacsa.2023.0140813.

[4] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, "Fine-Tuned IndoBERT Based Model and Data Augmentation for Indonesian Language Paraphrase Identification," *Rev. Intell. Artif.*, vol. 37, no. 3, pp. 733–743, Jun. 2023, doi: 10.18280/ria.370322.

[5] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Electr. Comput. Eng. IJECE*, vol. 13, no. 5, p. 5641, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[6] S. Aras, M. Yusuf, R. Y. Ruimassa, E. A. B. Wambrauw, and E. B. Pala'langan, "Sentiment Analysis on Shopee Product Reviews Using IndoBERT," *J. Inf. Syst. Inform.*, vol. 6, no. 3, pp. 1616–1627, Sep. 2024, doi: 10.51519/journalisi.v6i3.814.

[7] F. Iscus and A. S. Girsang, "Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method," *Int. J. Eng. Trends Technol.*, vol. 70, no. 12, pp. 281–288, Dec. 2022, doi: 10.14445/22315381/ijett-v70i12p226.

[8] Ibadurrohman Irfan Fatani, "Twitter, Instagram, Youtube Speak: Understanding Sentiments on LRT Jabodebek Services via Inset Lexicon, IndoBERT and BERTopic Approaches," *J. Electr. Syst.*, vol. 20, no. 4s, pp. 1028–1035, Apr. 2024, doi: 10.52783/jes.2147.

[9] Y. A. Singgalen, "Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data," vol. 6, no. 2, 2025.

[10] S. Pecar, M. Simko, and M. Bielikova, "Sentiment Analysis of Customer Reviews : Impact of Text Pre-processing," *2018 World Symp. Digit. Intell. Syst. Mach. DISA*, pp. 251–256, 2018, doi: 10.1109/DISA.2018.8490619.

[11] R. K. Bania, "COVID-19 Public Tweets Sentiment Analysis using TF-IDF and Inductive Learning Models Handwritten Assamese Character Recognition using Texture and Diagonal Orientation features with Artificial Neural Network View project COVID-19 Public Tweets Sentiment An," no. December, 2020, [Online]. Available: https://www.researchgate.net/publication/346572350

[12] D. R. Firmansyah and E. Lestariningsih, "Analisis Sentimen Ulasan Aplikasi Smart Campus Unisbank di Google Playstore Menggunakan Algoritma Naive Bayes," *J. JTIK J. Teknol. Inf. Dan Komun.*, vol. 8, no. 2, pp. 498–507, Apr. 2024, doi: 10.35870/jtik.v8i2.1882.

[13] D. Marutho, M. Muljono, R. Supriadi, and P. Purwanto, "Optimizing Aspect Term Extraction and Sentiment Classification through Attention Mechanism and Sparse Attention Techniques," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, pp. 1004–1015, Oct. 2024, doi: 10.22266/ijies2024.1031.75.

[14] Primanda Sayarizki, Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Indones. J. Comput. Indo-JC*, vol. Vol. 9 No. 2, pp. 61-72 Pages, Aug. 2024, doi: 10.34818/INDOJC.2024.9.2.934.

[15] E. Arif, S. Suherman, and A. P. Widodo, "Predicting Stock Prices of Digital Banks: A Machine Learning Approach Combining Historical Data and Social Media Sentiment from X," *Ingénierie Systèmes Inf.*, vol. 30, no. 3, Mar. 2025, doi: 10.18280/isi.300313.

[16] M. Rosidin, M. F. Gustafi, and S. A. Pratiwi, "Optimizing nazief adriani's stemmer algorithm in detecting indonesian word errors using sastrawi," no. 3.