Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

Improving Sentiment Classification of Kredit Pintar Reviews Using IndoBERT, SMOTE, and Stacking Ensemble

Ayu Safitri[1], Muhammad Risaldi[2], Muh Naufal Ramadhani Alwi[3], Dewi Fatmarani Surianto*[4], Nur Fadilah[5], Jumadi M Parenreng[6]

[1,2,3,4,5,6]Computer Engineering, Faculty of Engineering, Universitas Negeri Makassar, Indonesia

Email: [4]dewifatmaranis@unm.ac.id

Received : Sep 27, 2025; Revised : Jan 11, 2026; Accepted : Jan 19, 2026; Published : Jul 15, 2026

Abstract

Kredit Pintar is one of the most widely used fintech applications in Indonesia, generating millions of user reviews on the Google Play Store that reflect diverse user experiences. These reviews provide valuable insights into application performance; however, extracting sentiment from such unstructured and imbalanced textual data remains a challenging task. This study aims to improve sentiment classification of Kredit Pintar user reviews by proposing a hybrid approach that integrates IndoBERT, SMOTE (Synthetic Minority Over-Sampling Technique), and a stacking ensemble model. From 2020 to 2024, 2,278 user reviews were classified into positive, neutral, and negative categories based on star ratings. SMOTE was employed to rectify class imbalance, whereas IndoBERT gathered contextual representations of the Indonesian language. Furthermore, a stacking ensemble combining IndoBERT, Random Forest, and SVM (Support Vector Machine) was implemented to enhance classification performance. Experimental results show that IndoBERT without data balancing achieved an accuracy of 84%, whereas the proposed combination of IndoBERT, SMOTE, and stacking ensemble consistently produced superior performance, achieving 92% accuracy, precision, recall, and F1-score. The findings demonstrate that integrating language-specific transformer models with data balancing and ensemble techniques effectively improves sentiment classification. This study contributes to the advancement of Indonesian-language natural language processing in the fintech domain and provides practical insights for fintech developers in understanding user perceptions and improving digital financial services.

Keywords: *IndoBERT, Kredit Pintar, Online Loan, SMOTE, Stacking Ensemble.*

1. INTRODUCTION

The swift progression of communication and information technology has impacted numerous facets of existence, including financial, necessitating enhanced comprehension to avert prospective economic issues [1]. One of the most prominent developments is online lending, which is increasingly recognized as a fast and convenient financial service for meeting community needs. Online lending refers to borrowing transactions conducted via digital platforms, where borrowers are required to pay interest in accordance with the agreed terms [2]. To regulate this practice, the Financial Services Authority (OJK) issued Pear-to-Pear (P2P) lending regulation Number 77 / PJOK.01 / 2016, which allows the public to access funds more practically through online platforms, thereby increasing public interest in this service [3]. However, a report from OJK revealed that complaints related to online lending increased by 30% in 2022, with many of the issues centered on high interest rates and inappropriate debt collection practices [4]. Although online lending makes borrowing transactions easier with just internet access [5]. This convenience also bring risks, including fraud, data theft, and aggressive billing practices that may harm users[6].

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

These infractions have prompted users to provide feedback on the application platform. Kredit Pintar is one of the most prevalent online loan application platforms in Indonesia [7]. Kredit Pintar is a fintech company that provides convenient online loan services [8]. The application has been downloaded more than 10 million times, with a rating of 4.3 and over 2 million reviews on the Google Play Store. These reviews consist of public comments and evaluations expressed through star ratings (ranging from one to five) accompanied by text feedback. User ratings and reviews are often considered as a reference by potential new users. Consequently, it is essential to further categorize the substance of these reviews to ascertain if they convey positive, negative, or neutral thoughts [9].

Prior research has examined the application of machine learning techniques for sentiment analysis within financial applications and digital platforms. A range of traditional machine learning techniques, including Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (K-NN), and Naïve Bayes, has been widely applied to sentiment classification tasks using user reviews collected from the Google Play Store and social media platforms. Empirical evidence from prior studies indicates that these approaches are capable of effectively distinguishing positive, neutral, and negative sentiments in online lending services. In particular, SVM and Random Forest are frequently adopted as the main classification models and have been reported to achieve reliable performance in applications such as Kredit Pintar, Kredivo, and UangMe [10], [11]. Furthermore, ensemble approaches such as Stacking Ensemble have been introduced to improve performance on imbalanced datasets, with results showing improved model accuracy and stability [12]. These findings suggest that combining multiple models can provide advantages over using a single model.

Conversely, research comparing several classical algorithms indicates that their effectiveness is strongly influenced by preprocessing quality and susceptibility to data noise, particular for models such aas SVM, Naïve Bayes, and K-NN [13], [14]. Hyperparameter optimization in Naive Bayes models can improve accuracy, but this approach tends to be dependent on dataset characteristics and requires intensive tuning [15]. Subsequent research has utilized modern transformer-based NLP techniques to address these limitations. For example, Latifah demonstrated the potential of IndoBERT in SMS spam classification, achieving 93% accuracy after applying Easy Data Augmentation (EDA). The combination of IndoBERT and EDA proved effective, albeit requiring significant computational resources [16].

While these approaches have demonstrated reasonable performance, they often rely on bag-of-words or TF-IDF representations, which are limited in capturing contextual meaning and linguistic nuances, particularly in Indonesian-language user reviews that often contain slang, informal expressions, and emotional cues. Furthermore, some studies still face challenges related to class imbalance, particularly in the neutral sentiment category, and often require complex optimization strategies such as extensive hyperparameter tuning or data augmentation, which can increase model complexity without guaranteeing consistent performance improvements.

This study develops a hybrid sentiment analysis framework by integrating IndoBERT woth SMOTE based resampling namely, the Synthetic Minority Over-Sampling Technique and a stacking ensemble model to mitigate the identified limitations. A key contribution of this work is the integration of language-specific transformer models with data balancing and ensemble learning strategies, aimed at improving sentiment classification in Indonesian fintech user reviews. IndoBERT is leveraged to capture bidirectional contextual representations of the Indonesian language, enabling a deeper understanding of semantic and emotional nuances [17], while SMOTE is applied to effectively mitigate class imbalance without relying on complex data augmentation methods [3]. Furthermore, the ensemble stacking model combines the strengths of IndoBERT, Random Forest, and Support Vector Machine (SVM) to improve the robustness and generalization of the classification. This study makes academic contributions by demonstrating an effective hybrid approach for Indonesian language sentiment analysis and practical

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

contributions by providing insights that can support fintech developers and policymakers in understanding user perceptions and improving digital financial services.

2. METHOD

The method of study consists of a series of processes designed to provide an efficient research process and achieve the defined objectives. The phases of this sentiment analysis study are illustrated in Figure 1.

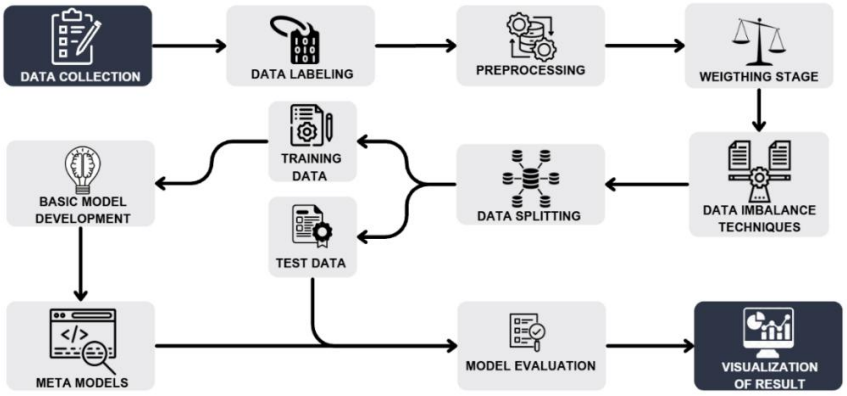

Figure 1. Research Workflow of the Proposed Sentiment Analysis Approach

2.1. Data Collection

The research data was sourced from user reviews of the Kredit Pintar application on the Google Play Store, with a total of 2,278 reviews published from 2020 to 2024. The data collection procedure utilized an automated web crawling technique implemented in Python, employing the google-play-scraper package within the Visual Studio Code environment.

Table 1. Sample Data

No.	Comment
1	Kecewa banget 😭😭😭 di saat perlu banget mau pinjam di tolak terus, suru tunggu berapa bulan lagi baru ajukan, setelah ajukan di tolak lagi 😭😭 udah 2x pinjam di sini pertama dan kedua aman2 cepat prosesnya setelah mau ke tiga di tolak terus berx x,.. kredit pintar tolong kasih penjelasanya 😭😭.
2	sangat membatu mudah dan fleksibel persyaratannya jga gk rumit limit lumayan lebih mudahnya saat membayar angsuran itu bisa di perpanjang jadi sangat membantu sekali. thank kredit pintar sdh membantu saat mendesak dan langsung cair ★★★★★★★★★★
...	.....
2278	Saya batu mendaftar.. Setelah pengisian data. Kok ada tulisan pengajuan sedang di verifikasi untuk di cairkan..Itu gimana maksdnya

The crawling technique involved collecting all accessible reviews from the past five years without implementing content-based exclusion criteria, hence ensuring comprehensive capture of user opinion. Each dataset has significant features, including star ratings, review content, and publishing timestamps.

This method facilitates rapid data gathering and guarantees the data's appropriateness for sentiment analysis based on sentiment categories obtained from user ratings [18]. The gathered sample data can be seen in Table 1.

2.2. Data Labeling

The labeling process was performed based on user rating scores obtained from the Kredit Pintar application on the Google Play Store. Sentiment label were divided into three categories: positive, neutral, and negative. User reviews with ratings for four and five stars were assigned a positive label, three-star reviews were considered neutral, while reviews rated dataset consists of 974 positive reviews, 124 neutral reviews, and 1,180 negative reviews. Representative sample of the labelled review data are shown in Table 2.

Table 2. Overview of the Rating-based Data Labeling Process

Label	Comment
Positive	Kredit pintar sangat membantu saya dalam keperluan yang mendadak, pengajuannya cukup mudah dan prosesnya cepat, rekomendasi terbaik.
Neutral	Saya ada kendala di pembayaran saat ini no rek virtual yang tertera tidak valid. Apa sulusi y terima kasih
Negative	Kurang fleksibel dan bunga terlalu tinggi di banding yang lain. Pencarian cepat adalah nilai plus.

2.3. Preprocessing

Data preprocessing is performed to prepare raw user reviews for use in sentiment analysis by reducing noise and ensuring text consistency [19]. Preprocessing steps include text normalization, data cleaning, handling of blank values, and case folding. During the normalization stage, non-standard and informal words commonly used in user reviews are manually converted into standard Indonesian. The cleaning process removes punctuation, emoticons, and special characters, leaving only letters, numbers, and spaces. If blank values are present, they are replaced with empty strings. Next, case folding is applied to convert all characters to lowercase to maintain uniform text representation in the dataset.

2.4. Weighting Stage

At this stage, calculations are performed to determine the value or weight of the extracted words. This method utilizes TF-IDF weighting namely, Term Frequency-Inverse Document Frequency to present word importance across documents. The frequency of a term's occurrence influences its weight assessment [20]. The formula used to calculate TF-IDF is presented in Formula 1:

$$TF - IDF_{t.d} = TF_{t.d} \times IDF_t \quad (1)$$

Description:

$TF\text{-}IDF_{t.d}$ = weight form of word t in one document

$TF_{t.d}$ = frequency of occurrence of word t in the document

IDF_t = inverse document frequency

$$IDF_t = \log\left(\frac{N}{DF_t + 1}\right) \quad (2)$$

Description:

N = number of documents

DF_t = number of documents containing word t

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

2.5. Data Imbalance Technique

The data imbalance technique is used to address situations where one class has significantly more data than others [21]. In this research, several techniques were applied to overcome data imbalance, namely SMOTE, Synonym EDA, ADASYN, and Random Oversampling. SMOTE mitigates class imbalance by generating synthetic data [22]. In sentiment analysis with three classes (positive, neutral, and negative), the dataset is often imbalanced because user reviews tend to be polarized into positive or negative, making neutral data the minority class and sometimes far smaller than the other classes [10]. Utilizing SMOTE enhances the sample size of the minority class while preserving the inherent attributes of the data, thereby augmenting the efficacy of the machine learning model [20].

Synonym EDA (Easy Data Augmentation) is a text augmentation technique that introduces variety into the dataset to enhance the model's text representation ability [23]. ADASYN, a method similar to SMOTE, generates additional synthetic sample with a stronger focus on the minority class [24]. Random oversampling equilibrates the dataset by replicating instances from the minority class, hence augmenting its size. augment the data in the minority class to get a twofold increase in the minority class data [25].

2.6. Data Splitting

At this stage, the dataset is partitioned into training and testing subsets, with 80% used for model training and the remaining 20% rederved for performance evaluation. This division follows previous studies that applied the 80:20 ratio and achieved good accuracy results [26].

2.7. Basic Model Development

This research employs IndoBERT, a transformer-based language model specifically pre-trained and tailored for the Indonesian language, as the primary model for text representation.

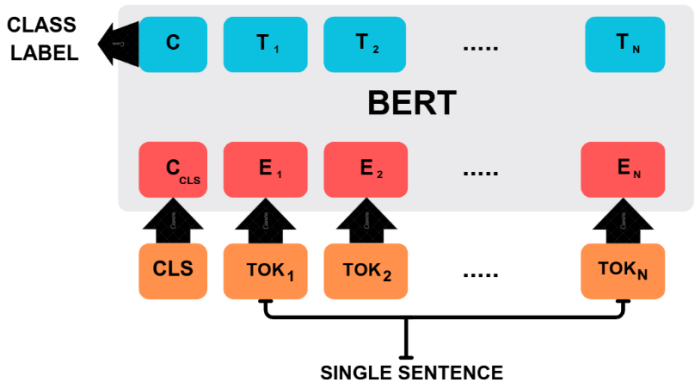

Figure 2. BERT Architecture

The pre-trained IndoBERT model underwent supervised fine-tuning for sentiment classification utilizing tagged user review data. During this fine-tuning phase, all model parameters were adjusted to synchronize the resultant language representations with the context of sentiment analysis. The fine-tuning configuration comprised three training epochs with a batch size of five datasets each iteration. The learning rate was established at 5 × 10^-5 to guarantee training stability. The training process was confined to 500 steps, with model performance assessed every 10 steps. The dataset was partitioned via a hold-out validation approach, with 80% allocated for training and the remaining 20% designated for evaluation. This design was selected to attain a compromise between training efficiency and model performance while preventing overfitting. An overview of the BERT architecture is presented in Figure 2.

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

2.8. Model Evaluation

The effectiveness of the trained model is in the evaluation phase. Model performance is evaluated using data excluded from the training phase. The model utilizes this unseen data to execute sentiment prediction and generates raw prediction results prior to the application of the activation function in the computation such as accuracy, precision, recall, F1-score, along with confusion matrix evaluation. The predicted values are subsequently transformed into probabilities ranging from 0 to 1 for each class.

2.9. Final Classification

The final classification stage involves the creation of a meta-model that amalgamates the forecasts of many base models to yield a more resilient final prediction using stacking-based ensemble learning. Ensemble stacking is utilized as the ultimate estimation approach to enhance overall model performance by capitalizing on the complementing advantages of each base model. In this approach, multiple base learners first produce individual outputs, which are subsequently integrated through a meta-model to obtain the final prediction. This study used three foundational models: IndoBERT, which generates probability-based predictions as the primary model, and Random Forest and SVC (Support Vector Classifier), both utilizing TF-IDF features derived from the training data. Logistic Regression serves as the ultimate estimator (meta-model), tasked with producing the final prediction derived from the outputs of the base models. The overall architecture of the stacking ensemble, including the integration of 5-fold cross-validation to reduce the risk of overfitting, is illustrated in Figure 3.

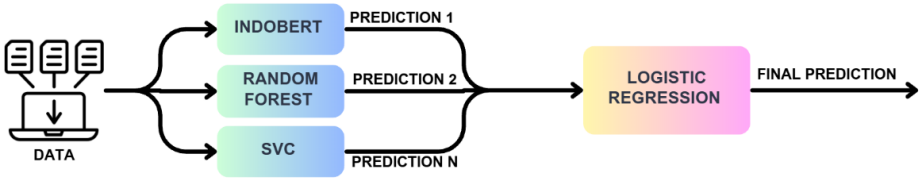

Figure 3. Stacking Ensamble Model

3. RESULT

This study involved data collecting using web crawling, yielding a total of 2,278 data points. The data were sourced from user reviews of the Kredit Pintar application in the Google Play Store, spanning the years 2020 to 2024. Consequently, data tagging was conducted according to the user ratings of the application. Reviews rated 4 and 5 were categorized as positive, a rating of 3 as neutral, and ratings of 1 and 2 as negative. This study's dataset comprises 974 good reviews, 124 neutral reviews, and 1,180 bad reviews. During the preprocessing phase, multiple procedures were executed, encompassing normalization, data cleansing, management of NaN values, and case folding through the conversion of text to lowercase. The preprocessing result are presented in Table 3.

Table 3. Summary of the Applied Text Preprocessing Steps and Results

Preprocessing	Comment
Initial Data	Aplikasi jelek bgt proses lama padahal data udah lengkap tp g di acc, tolong data saya jangan disalahgunakan hapus data² dan akun saya.
Normalization and cleaning	Aplikasi jelek banget proses lama padahal data sudah lengkap tapi tidak di acc tolong data saya jangan disalahgunakan hapus data dan akun saya
Case Folding (lower)	aplikasi jelek banget proses lama padahal data sudah lengkap tapi tidak di acc tolong data saya jangan disalahgunakan hapus data dan akun saya

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

Table 3 presents the outcomes of the data preprocessing phase. Following these results, the subsequent step is to compute the TF-IDF weights utilizing a specified formula. The result of the TF-IDF weighting process are presented in Table 4.

Table 4. Computed TF-IDF Term Weights for the Review Dataset

Term	TF1	TF2	TF3	DF	IDF+1	TF-IDF 1	TF-IDF 2	TF-IDF 3
Indonesia	0	0	0	8	6.651	0	0	0
Limit	0	0	0.111	393	2.757	0	0	0.307
Lumayan	0	0	0.192	52	4.779	0	0	0.922
Masalah	0	0.116	0	60	4.6367	0	0.539	0
Masih	0.123	0	0	208	3.393	0.419	0	0
Melalui	0.188	0	0	34	5.204	0.983	0	0

Table 4 presents an example of word weighting results using the TF-IDF method. This table shows how several terms are represented in different documents through their Term Frequency (TF), Document Frequency (DF), Inverse Document Frequency (IDF), and final Term Frequency – Inverse Document Frequency (TF-IDF) values. Terms that appear in fewer documents, such as "lalu" and "lumayan," have higher IDF values, resulting in higher TF-IDF weights for certain documents. This representation demonstrates that the TF-IDF method is able to suppress more informative words and reduce the influence of frequently occurring words, thus supporting the formation of effective text features before the sentiment classification process is carried out.

3.1. Comparison of Data Balancing Techniques

Data imbalance management is executed to mitigate the prevalence of the majority class and enhance model efficacy on the minority class. This experiment employs four data balancing techniques: SMOTE, EDA Synonym, ADASYN, and Random Oversampling. Each strategy was evaluated under identical experimental conditions, using the same dataset split and an identical IndoBERT configuration to ensure a fair comparison. The experimental results for each approach are presented in Table 5.

Table 5. Performance Comparison of Model using Different Imbalance Techniques

Model	Accuracy	Precision	Recall	F1-Score
SMOTE	**0.90**	**0.91**	**0.90**	**0.90**
EDA Synonyms	0.89	0.89	0.89	0.89
ADASYN	0.89	0.89	0.88	0.89
Random Oversampling	0.91	0.90	0.91	0.90

Based on Table 5, the SMOTE and Random Oversampling techniques demonstrated higher performance than EDA Synonym and ADASYN. Although Random Oversampling produced slightly higher accuracy (0.91) than SMOTE (0.90), Random Oversampling operates by replicating instances in the minority class, which may elevate the likelihood of overfitting and diminish the model's generalizability. In contrast, SMOTE generates new synthetic data through feature space interpolation, allowing the model to learn more diverse decision boundaries. Therefore, based on quantitative results and statistical evaluations, SMOTE was selected as the data balancing technique for further experiments. The comparison of data quantities across classes before and after applying each data balancing technique is illustrated in Figure 4.

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

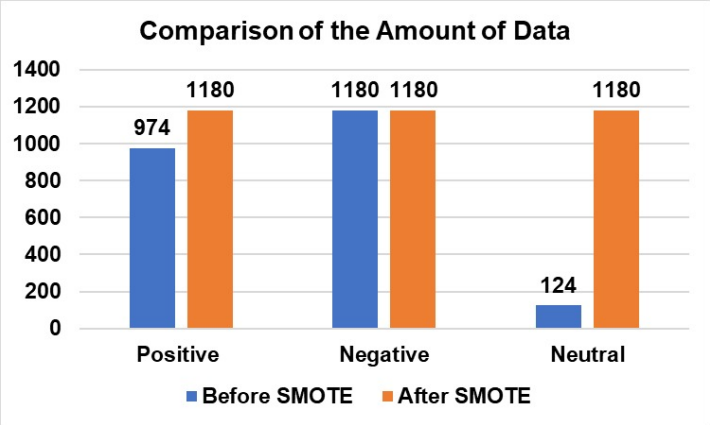

Figure 4. Comparison Graph of the Amount of Data

As seen in Figure 4, the comparison of the amount of data before and after SMOTE was successfully performed, balancing data. The data before SMOTE amounted to 974 positive, 1180 negative, and 124 neutral. Following SMOTE application, the dataset achieved class balance, with 1,180 instances in each of the positive, negative, and neutral categories. The dataset is partitioned into training and testing sets, with 80% allocated for model learning and the remaining 20% reserved for performance model. This divide yields 2,832 data for training and 708 for testing, which will be employed for TF-IDF feature extraction and the IndoBERT model. The TF-IDF feature will be used in the model Random Forest and SVM models, which are the basic model used in the meta-model, while the IndoBERT model is the basic model built to obtain probability values as input later in the meta-model.

3.2. Comparison of IndoBERT Performance with and without SMOTE

The first experiment by testing the IndoBERT model without using SMOTE gave less than optimal results. Without SMOTE, the tested data was unbalanced, with each neutral data amounting to 124, positive data amounting to 974, and negative data amounting to 1180. From this data, 20% was taken as test data where the number of data for each class was 32 for neutral, 228 for negative, and 196 for positive. This data was used in the evaluation stage of the classification model, which resulted in a performance comparison between the IndoBERT model using SMOTE and without SMOTE. The confusion matrix based comparison is shown in Figure 5.

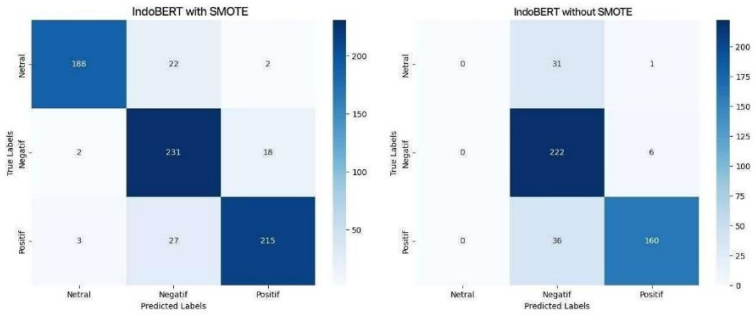

Figure 5. Confusion Matrix Comparison IndoBERT Model using SMOTE and WIthout SMOTE

Figure 5 shows a confusion matrix analysis showing that the IndoBERT model without SMOTE was unable to correctly classify any data in the neutral class, with most neutral data incorrectly classified into the negative or positive classes. This finding suggests that the model is biassed toward the dominant class and exhibits limited generalization for underrepresented classes. By incorporating SMOTE, the model demonstrates improved recognition of the neutral class while simultaneously reducing

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

misclassification across all sentiment categories. The results affirm that data balancing by SMOTE is essential for enhancing IndoBERT's generalization abilities, especially in identifying minority sentiment classes. Performance metric of the indoBERT model presented in Table 6.

Table 6 shows the difference in results between the IndoBERT model using SMOTE and without SMOTE. The performance accuracy of the IndoBERT model using SMOTE reached 0.90, indicating that the SMOTE technique provided a major change to the minority class, making the model better at understanding and analyzing data even though there was still data that was not successfully classified in its class. The neutral class successfully classified 188 data in its class and misclassified 24 data. The negative class successfully classified 231 data in its class and misclassified 20 data. The positive class successfully classified 215 data in its class and misclassified 30 data. Meanwhile, the performance accuracy of the IndoBERT model without SMOTE decreased by 0.6, resulting in 0.84 performance accuracy.

Table 6. Comparison of IndoBERT Performance using SMOTE and Without SMOTE

Model	Accuracy	Precision	Recall	F1-Score
With SMOTE	0.90	0.97	0.89	0.93
		0.82	0.92	0.87
		0.91	0.88	0.90
Without SMOTE	0.84	0	0	0
		0.77	0.97	0.86
		0.96	0.82	0.88

The positive class successfully read 160 data in its class, and as many as 36 were misclassified as negative classes. In the negative class, 222 data in its class were successfully read; misclassified. as many as 6 were read as the positive class. In the neutral class, no data was successfully read in its class. As many as 31 data were read in the negative class and 1 data was read in the positive class. The results indicate that the model exhibited overfitting on the majority class words, resulting in diminished comprehension of the minority class words. Therefore, the SMOTE technique is the choice for the technique imbalance data to be combined with the IndoBERT model.

3.3. Meta-Model with Stacking Ensemble

The combination of SMOTE and IndoBERT has achieved promising results in improving data understanding. However, the model still produces classification errors in its predictions. To address this, a classification test was conducted using a meta-model with a stacking ensemble approach, in which the output probability distributions generated by IndoBERT are subsequently fed into the meta-model. The results of confusion matrix comparison are illustrated in Figure 6.

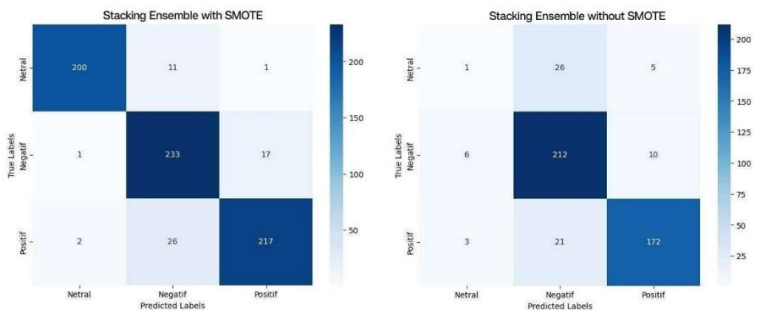

Figure 6. Confusion Matrix Comparison Stacking Model using SMOTE and Without SMOTE

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

Figure 6 presents a confusion matrix study indicating that the meta-model utilizing the ensemble stacking method enhances model performance. The neutral class accurately predicted 200 data points, while 12 data points were predicted inaccurately. The negative class accurately predicted 233 data points, with 18 misclassifications, whereas the positive class properly predicted 217 data points and misclassified 28 data points. The subsequent experiment, evaluating the Stacking Ensemble using data not subjected to the SMOTE approach, demonstrates that the model necessitates balanced data to achieve favorable outcomes in sentiment prediction. In the neutral class, just one data point was accurately predicted, whilst 31 data points were misclassified. In the negative class, it accurately predicted 212 data points, whereas 16 data points were projected wrongly. Simultaneously, the positive class accurately predicted 172 data points, while 24 data points were misclassified. The performance comparison of the stacking ensemble model is presented in Table 7.

Table 7. Comparison of Stacking Model Performance using SMOTE and Without SMOTE

Model	Accuracy	Precision	Recall	F1-Score
With SMOTE	0.92	0.97	0.96	0.96
		0.87	0.91	0.89
		0.92	0.88	0.90
Without SMOTE	0.84	0.10	0.3	0.5
		0.82	0.93	0.87
		0.92	0.88	0.90

Table 7 indicates that the Stacking Ensemble attained an accuracy of 0.92, reflecting a 0.2 improvement over the preceding model. The Stacking Ensemble utilizing SMOTE shown commendable efficacy in prediction accuracy. In contrast, the stacking ensemble without SMOTE shows no notable performance improvement, achieving an accuracy of 0.84. Most classification errors originate from the neutral class, reflecting the absence of an explicit mechanism to address class imbalance. Through the execution of multiple experimental situations, the optimal results can be evaluated.

3.4. Comparison of All Experimental Scenarios

Following the execution of multiple experimental situations, the performance of each model is illustrated by the variance in accuracy levels, culminating in the F1-score, as shown in Table 8.

Table 8. Comparison of All Experiment Scenario

Experiment	Accuracy	Precision	Recall	F1-score
IndoBERT	0.84	0.58	0.60	0.58
IndoBERT + SMOTE	0.90	0.90	0.89	0.90
IndoBERT + Stacking	0.84	0.61	0.61	0.61
IndoBERT + SMOTE + Stacking	**0.92**	**0.92**	**0.92**	**0.92**

Table 8 shows the optimal performance outcomes derived from diverse trials utilizing distinct techniques and methodologies. The IndoBERT model, absent the SMOTE approach, attained an accuracy of 0.84, whereas the incorporation of SMOTE enhanced performance to 0.90. Moreover, IndoBERT models that addressed data imbalance and were augmented with the stacking model exhibited increased classification performance, attaining an average accuracy of 0.92. The IndoBERT model, when applied with stacking but neglecting data imbalance, achieved a lower accuracy of 0.84. This discovery highlights the necessity of addressing data imbalance in sentiment analysis, since it

Vol. 7, No. 3, June 2026, Page. 2411-2424
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5342

substantially affects model efficacy. Nevertheless, the IndoBERT model, despite the absence of data balancing strategies, has notable proficiency in sentiment categorization. A comparative analysis of classification models using accuracy and F1-score is shown in Figure 7.

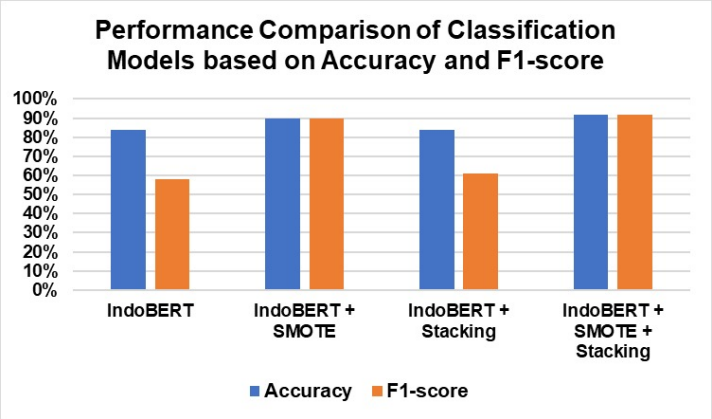

Figure 7. Comparison of Accuracy and F1-score Across Classification Models

Figure 7 presents a graphical comparison of the four models with respect to their accuracy and F1-score. The baseline IndoBERT model has relatively high accuracy, but its lower F1-score indicates weakness in handling class imbalance. The application of SMOTE significantly improves both metrics, confirming its effectiveness in balancing sentiment data. The stacking ensemble model without SMOTE provides moderate improvements, especially in accuracy, but still shows a relatively low F1-score. The optimal performance is achieved with the amalgamation of IndoBERT, SMOTE, and Stacking, resulting in elevated and balanced accuracy and F1-score, so illustrating that the synthesis of data balancing and ensemble learning methodologies may provide a more resilient sentiment classification model.

3.5. Word Cloud Visualization for Sentiment Theme

Figure 8. Word Cloud Visualization for Sentiment Theme

The word cloud picture depicts the frequency of the most prevalent terms in user evaluations of Kredit Pintar, categorized into positive, negative, and neutral sentiments. This word cloud visualization are illustrated in Figure 8.

Figure 8 shows the frequency distribution of word in positive, negative, and neutral reviews. Positive sentiment dominated by words like “cepat”, “mudah”, dan “bantu” indicating that users felt supported by the quick loan application and disbursement process and the app's ease of use. Conversely, negative sentiment dominated by words like “tagihan”, “telat”, dan “kecewa” reflecting user complaints about late payments, the billing system, and the service experience. Meanwhile, the neutral sentiment word cloud featured a high number of informative words like “pengajuan”, “kredit”, dan “aplikasi” indicating that reviews focused more on describing application usage without clearly expressing positive or negative emotions.

4. DISCUSSIONS

This study demonstrates that the integration of IndoBERT, SMOTE, and stacking ensembles yields superior performance compared to previous approaches in sentiment analysis of fintech applications. The proposed model achieves 92% accuracy, surpassing various conventional machine learning-based methods, which generally range from 79% to 90%. Support Vector Machine and Random Forest-based approaches in previous studies generally yield accuracy below 85%, while ensemble methods without a transformer model achieve performance of around 87%. Although hyperparameter optimization in Multinomial Naive Bayes can improve accuracy to nearly 91%, this result is still below the performance of the proposed model. This comparison indicates that utilizing an Indonesian-language transformer model combined with data balancing techniques and ensemble learning can provide significant and more consistent performance improvements.

The enhanced model performance is mostly because to IndoBERT's capacity to comprehend the context and subtleties of Indonesian, encompassing informal language, slang, and emotive expressions commonly present in user evaluations. Unlike TF-IDF-based approaches, IndoBERT uses bidirectional contextual embedding, enabling a deeper understanding of sentence meaning. Furthermore, the application of SMOTE successfully addressed data imbalance, particularly in the neutral sentiment class, which failed to be correctly classified when SMOTE was not used. Without SMOTE, model accuracy dropped to 84%. The use of a stacking ensemble further enhanced the model's robustness by combining the advantages of IndoBERT, Random Forest, and SVM.

This research experimentally demonstrates that the integration of an Indonesian transformer model with data balancing and ensemble learning approaches significantly enhances sentiment classification performance for tasks invoving Indonesian natural language processing and related informatics applications. The results of this study emphasize the importance of contextual language modeling in languages with com plex morphological structures such as Indonesian and extend previous research in ensemble-based sentiment analysis.

Practically, the results of this study provide benefits to fintech application developers and policymakers. More accurate sentiment classification can help understand user satisfaction and complaints, support service improvements, and strengthen consumer protection. This approach also has potential applications in other contexts such as fraud detection, reputation monitoring, and customer behavior analysis.

This study has several limitations. First, the data used is only from the Google Play Store, so generalization to other platforms is limited. Second, although SMOTE improves the performance of the minority class, misclassification still occurs in ambiguous reviews. Furthermore, IndoBERT's computational requirements are relatively high compared to simpler methods, making it less ideal for systems with limited resources. Future research could explore lighter hybrid architectures, expand data sources to multiple platforms, and apply cross-domain approaches to improve model generalization.

5. CONCLUSION

This research illustrates that the amalgamation of IndoBERT, SMOTE, and stacking ensemble techniques significantly enhances sentiment analysis efficacy in Indonesian finance applications, specifically Kredit Pintar. By mitigating data imbalance, the proposed model achieved strong performance, with accuracy, precision, recall, and F1-score all 92%. The finding demonstrate that the research goals were achieved and confirm the effectiveness of combining the Indonesian transformer model with data balancing and ensemble learning techniques. This technique assists fintech developers in effectively interpreting customer comments, while also contributing academically to the advancement of sentiment analysis and Indonesian natural language processing.

Despite its contributions, this study has certain limitations that should be considered when interpreting the results. As the study relies solely on Google Play Store data, its findings may not fully generalize to other app distribution platforms. Neutral sentiment remains a challenge due to its ambiguous nature, even with the use of SMOTE. Furthermore, IndoBERT's computational complexity can be a constraint in resource-constrained environments. Future research could explore the use of lighter architectures, expand data sources to multiple platforms, and involve linguistic experts in the labeling process. Overall, this study confirms IndoBERT's strategic role in the development of sentiment analysis in informatics, particularly for Indonesian fintech applications.

CONFLICT OF INTEREST

The authors report that they have no financial or personal conflicts of interest associated with this publication.

REFERENCES

[1] W. Nopriansyah and N. S. Wafi, "Literasi Keuangan Digital: Bahaya dan Dampak Pinjaman Online Ilegal Bagi Mahasiswa," *AKM: Aksi Kepada Masyarakat*, vol. 5, no. 1, pp. 421–432, 2024, doi: 10.36908/akm.v5i1.1118.

[2] R. Afandi, M. Afdal, and R. Novita, "Analisis Sentimen Masyarakat Terhadap Pinjaman Online di Twitter Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest N eighbor," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 596–605, 2024, doi: 10.47065/bits.v6i2.5300.

[3] R. N. Ikhsani and F. F. Abdulloh, "Optimasi SVM dan Decision Tree Menggunakan SMOTE Untuk Mengklasifikasi Sentimen Masyarakat Mengenai Pinjaman Online," *Jurnal Media Informatika Budidarma*, vol. 7, no. 4, p. 1667, 2023, doi: 10.30865/mib.v7i4.6809.

[4] F. Farhan, F. Hamdani, N. L. V. P. Astuti, H. A. Haekal Fiqry, and M. R. Aulia, "Reformasi Hukum Perlindungan Data Pribadi Korban Pinjaman Online (Perbandingan Uni Eropa dan Malaysia)," *Indonesia Berdaya*, vol. 3, no. 3, pp. 567–576, 2022, doi: 10.47679/ib.2022264.

[5] F. Kurniawan, D. Suhariyanto, and Hartana, "Perlindungan Konsumen Terhadap Pinjaman Online Atas Penyebaran Data Pribadi," *INNOVATIVE: Journal Of Social Science Research.*, vol. 4, no. 1, pp. 2817–2829, 2024, doi: 10.31004/innovative.v4i1.7857.

[6] F. Novika, N. Septiavani, and I. M. Indra, "Pinjaman Online Ilegal Menjadi Bencana Sosial Bagi Generasi Milenial," *Management Studies and Entrepreneurship Journal*, vol. 3, no. 3, pp. 1174–1192, 2022, doi: doi.org/10.37385/msej.v3i3.857.

[7] D. A. K. Putra, "Karakteristik Verba Dan Adjektiva Dalam Iklan Aplikasi Pinjaman Online," *Adabiyyāt: Jurnal Bahasa dan Sastra*, vol. 6, no. 1, p. 42, 2022, doi: 10.14421/ajbs.2022.06103.

[8] Y. B. Putri, M. N. Hidayati, and N. Istiani, "Perlindungan Hukum Atas Klausula Baku yang Merugikan Debitur Pada Pinjaman Online Kredit Pintar," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, no. 3, pp. 16473–16487, 2024, doi: 10.31004/innovative.v4i3.12548.

[9] A. Agustin, S. Andrean, S. Susanti, R. Rahmiati, and H. Hamdani, "Review Aplikasi Kredivo Menggunakan Analisis Sentimen Dengan Algoritma Support Vector Machine," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 9, no. 1, pp. 39–49, 2023, doi: 10.36341/rabit.v9i1.4107.

[10] M. Iqbal, M. Afdal, and R. Novita, "Implementasi Algoritma Support Vector Machine Untuk Analisa Sentimen Data Ulasan Aplikasi Pinjaman Online di Google Play Store," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1244–1252, 2024, doi: 10.57152/malcom.v4i4.1435.

[11] Y. A. Wibisono, M. Afdal, M. Mustakim, and R. Novita, "Implementasi Algoritma Random Forest Untuk Analisa Sentimen Data Ulasan Aplikasi Pinjaman Online di Google Play Store," *Buiding of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, pp. 1244–1252, 2024, doi: 10.47065/bits.v6i2.5368.

[12] A. Munna and E. Zuliarso, "Interpretasi model Stacking Ensemble untuk analisis sentimen ulasan aplikasi pinjaman online menggunakan Lime," *AITI: Jurnal Teknologi Informasi*, vol. 21,

no. 2, pp. 183–196, 2024, doi: 10.24246/aiti.v21i2.183-196.

[13] R. S. Arischo and D. Damayanti, "Analisis Sentimen Pinjaman Online di Twitter dengan Metode Naive Bayes Classifier dan SVM," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, p. 1120, 2024, doi: 10.30865/mib.v8i2.7406.

[14] M. Husni, A dan Randi, "Analisis Sentimen Pengguna Aplikasi Kredivo Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Inovasi Global*, vol. 2, no. 3, pp. 543–551, 2024, doi: 10.58344/jig.v2i8.150.

[15] S. Nada Apsariny, Sediono, N. Chamidah, E. Ana, and A. Kurniawan, "Sentiment Analysis of User Reviews Based on Naïve Bayes," *Jurnal Ilmiah Indonesia*, vol. 7, no. 1, 2022, doi: 10.36418/syntax-literate.v7i1.6012.

[16] N. Latifah, R. Dwiyansaputra, and G. S. Nugraha, "Multiclass Text Classification of Indonesian Short Message Service (SMS) Spam using Deep Learning Method and Easy Data Augmentation," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 3, pp. 663–676, 2024, doi: 10.30812/matrik.v23i3.3835.

[17] R. Merdiansah and A. Ali Ridha, "Sentiment Analysis of Indonesian X Users Regarding Electric Vehicles Using IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 221–228, 2024, doi: 10.55338/jikomsi.v7i1.2895.

[18] A. Riskiyah, T. M. Fahrudin, and K. M. Hindrayani, "Analisis Sentimen Kepuasan Pelayanan Transportasi Online Gojek Menggunakan Algoritme Extreme Learning Machine," *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 1273–1285, 2024, doi: 10.46306/lb.v5i2.

[19] P. W. Rahayu, P. W. Gunawan, I. M. D. Ardiada, and N. P. M. A. Putri, "Analisis Sentimen Pada Media Sosial Twitter Terhadap Kepolisian Menggunakan Algoritma Support Vector Machine," *Jurnal Informasi Dan Komputer (JIK)*, vol. 12, no. 2, pp. 120–125, 2024, doi: 10.35959/jik.v12i02.546.

[20] S. Sasmita, R. N. Jariah S.Intam, D. F. Surianto, and M. F. B, "Analisis Sentimen Terhadap Kontroversi Putusan MK Mengenai Usia Capres-Cawapres Menggunakan Multi-Layer Perceptron Dengan Teknik SMOTE," *Faktor Exacta*, vol. 17, no. 2, p. 188, 2024, doi: 10.30998/faktorexacta.v17i2.22442.

[21] A. F. B. Sajiwo, B. Rahmat, and A. Junaidi, "Klasifikasi Indeks Standar Pencemaran Udaran (Ispu) Menggunakan Algoritma Xgboost Dengan Teknik Imbalanced Data (SMOTE)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4699.

[22] A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *International Journal of Information and Education Technology*, vol. 13, no. 2, pp. 289–295, 2023, doi: 10.18178/ijiet.2023.13.2.1806.

[23] I. A. Oktariansyah, F. R. Umbara, and F. Kasyidi, "Klasifikasi Sentimen Untuk Mengetahui Kecenderungan Politik Pengguna X Pada Calon Presiden Indonesia 2024 Menggunakan Metode IndoBert," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 636–648, 2024, doi: 10.47065/bits.v6i2.5435.

[24] M. T. triani B. Sirait, N. S. Fathonah, and M. N. Fauzan, "Pemanfaatan Algoritma ADASYN dan Support Vector Machine dalam Meningkatkan Akurasi Prediksi Kanker Paru-Paru," *JATI:Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 5, pp. 8773–8778, 2024, doi: 10.36040/jati.v8i5.10752.

[25] N. C. Nugraha, H. Hikmayanti, J. Indra, and A. R. Juwita, "Implementasi Metode Resampling Dalam Menangani Data Imbalance Pada Klasifikasi Multiclass Penyakit Thyroid," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 890–900, 2024, doi: 10.47065/bits.v6i2.5652.

[26] R. A. Azizah, F. Bachtiar, and S. Adinugroho, "Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 3, pp. 605–614, 2022, doi: 10.25126/jtiik.2022935751.