Journal of Soft Computing Exploration
Homepage: https://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Improved playstore review sentiment classification accuracy with stacking ensemble

Dwi Budi Santoso[1*], Aliyatul Munna[2], Dewi Handayani Untari Ningsih[3]
[1]Department of Information System, Universitas Stikubank, Indonesia
[2]Department of Master of Information Technology, Universitas Stikubank, Indonesia
[3]Department of Informatic Engineering, Universitas Stikubank, Indonesia

Article Info

Article history:

Received November 17, 2023
Revised March 17, 2024
Accepted March 18, 2024

Keywords:

Sentiment classification
Playstore review
Stacking ensemble
Logistic regression
Meta model

ABSTRACT

In today's digital era, user reviews on the Playstore platform are an invaluable source of information for developers, offering insights that are critical for service improvement. Previous research has explored the application of stacking ensemble methods, such as in the context of predicting depression among university students, to enhance prediction accuracy. However, these studies often do not explicitly detail the data acquisition process, leaving a gap in understanding the applicability of these methods to different domains. This research aims to bridge this gap by applying the stacking ensemble approach to improve the accuracy of sentiment classification in Playstore reviews, with a clear exposition of the data collection method. Utilizing Logistic Regression as the meta classifier, this methodology is executed in several stages. Initially, data was collected from user reviews of online loan applications on Google Playstore, ensuring transparency in the data acquisition process. The data is then classified using three basic models: Random Forest, Naive Bayes, and SVM. The outputs of these models serve as inputs to the Logistic Regression meta model. A comparison of each base model output with the meta model was subsequently carried out. The test results on the Playstore review dataset demonstrated an increase in accuracy, precision, recall, and F1 score compared to using a single model, achieving an accuracy of 87.05%, which surpasses Random Forest (85.6%), Naive Bayes (85.55%), and SVM (86.5%). This indicates the effectiveness of the stacking ensemble method in providing deeper and more accurate insights into user sentiment, overcoming the limitations of single models and previous research by explicitly addressing data acquisition methods.

Corresponding Author:

Dwi Budi Santoso,
Department of Information System,
Universitas Stikubank,
Jl. Tri Lomba Juang, Mugassari, Kota Semarang, Jawa Tengah 50241, Indonesia
Email: dbs@edu.unisbank.ac.id
https://doi.org/10.52465/joscex.v5i1.247

1. INTRODUCTION

The digital era has changed the way we interact with various services, including the financial sector such as online loan services. Platforms like Playstore are not only places to download applications but also forums where users share their opinions and experiences [1]. These reviews provide valuable data that reflects user perception and satisfaction, serving as direct feedback for app developers [2]. However, the large volume of reviews and diversity of user expressions make manual analysis impractical and require automation solutions.

Automating sentiment classification using machine learning methods has emerged as a powerful tool to effectively process and analyze these reviews [3]. Conventional machine learning models such as Random Forest, Naive Bayes, and Support Vector Machine (SVM) have been widely used, but each has certain limitations and biases in predicting sentiment. Random Forest can be used to classify the sentiment of e-commerce product reviews [4]. This model shows good capabilities in handling diverse and complex data, with fairly high accuracy. However, these models can overfit the training data and sometimes have difficulty generalizing to unseen data [5].

Sentiment analysis of social media posts using Naive Bayes has proven to be efficient and fast in training, but has limitations, but tends to have higher precision than recall [6]. The Support Vector Machine (SVM) algorithm is effective in classifying positive and negative sentiment, especially in data with high dimensions, but requires quite complicated parameter settings to avoid overfitting [7], [8].

In previous research on depression prediction among university students, various machine learning models were integrated using a stacking ensemble method. Subsequently, a comparison of evaluation metrics from each model was conducted. The results revealed that the stacking ensemble approach, combined with logistic regression, yielded the highest accuracy performance, achieving 91.22% without data balancing, and 94.69% with data balancing through the oversampling method. The implementation of the stacking ensemble technique, utilizing logistic regression, demonstrated superior performance when compared to the single-method Bayesian network, which attained an accuracy of 87.72% without any data balancing process [9]. Building upon previous studies, this research will utilize data from online loan applications, where data classification is conducted using single-method approaches such as Naive Bayes, SVM (Support Vector Machine), and Random Forest. Meanwhile, the stacking ensemble method is implemented utilizing Logistic Regression as the meta-classifier. This approach aims to leverage the complementary strengths of these diverse algorithms to enhance the predictive accuracy and reliability of sentiment analysis within the context of online financial services feedback.

2. METHOD

In this research, the dataset that has been collected and pre-processed will be used as input for sentiment classification in each base model, namely Random Forest, Naïve Bayes and SVM. The output from each base model is used as input to the Logistic Regression meta model. After that, the model was evaluated and analyzed and translated, as seen in Figure 1.

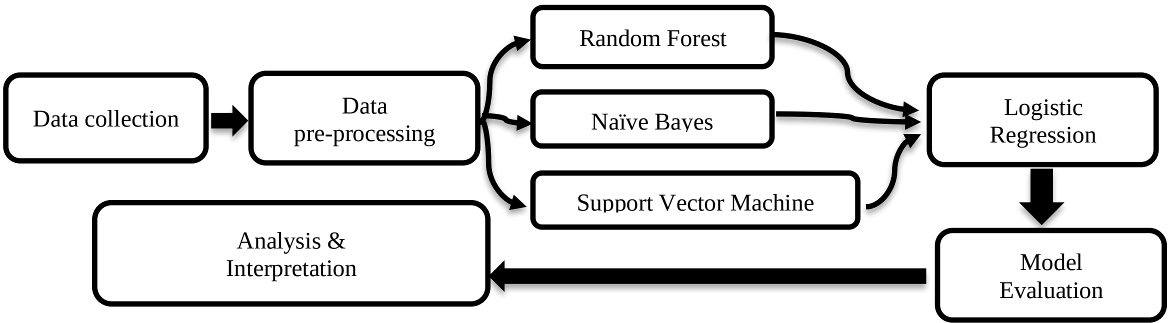

Figure 1. Research method

This research dataset consists of user reviews of online loan applications with the id com.adakami.dana.kredit.pinjaman available on Playstore. This review was collected using a scraping technique using the google-play-scraper python library. The data downloaded was 10,000 with a composition of 6,162 negative reviews with a score of 1 and 3,838 positive reviews with a score of 5. Table 1 shows sample data.

Table 1. Component classification

Content
Pengajuan saya sudsh ditolak 2x dan saya tidak pernah terlambat untuk membayar Apa alasan ada kami mengurangkan kredit score tanpa sepengetahuan saya limit saya jadi turun dan disuruh menambahkan informasi untuk meningkatkan limit Maksudnya apa?? Semau hati mengurangkan credit score saya 8x saya mengajukan 8x kali saya tidak pernah telat membayar Kebijakan apa yg terjadi dengan pihak ada kami??? Kalau tidak bisa membantu lg mohon data saya jangan disebar luaskan TRIMAKASIH
Aplikasi apaan ni? Dpt rekom kalo aplikasi ini cpat prosesnya, lgsg download malah gagal, trus disuruh perbaikin score credit goblok, baru juga mau minjem, ga jelas, kalo ga ada duit buat minjemin gausa buat aplikasi pinjol. Wkwkwk

Content
Pinjol bedebah…promo2 versi terbaru ketika install mau ngajukan ulang ..limit anda kurang…kgak usah sms….kgak usah promo ke saya….mirip orang kaya mau bagi2 rejeki niatnya cuma pamer….biar orang pada install.. Pinjol taik …pelanggaran kode etik…tutup aja nih aplikasi…OJK SWI mana suaranya…
Cukup bagus tapi tolong untk info penagihan bisa melalui notifikasi aplikasi saja atau di SMS, terlbh dahulu sebelum melakukan telepon? Sangat tidak sopan pagi2 telepon dengan nada bicara yang tidak menyenangkan
Tolong data saya hapus smuanya sya gamau data sya dipake yg ga bener soalnya aneh pas pertama verifikasi data biasa aja jalan normal ga gangguan,, knpa pas mau mengajukan pinjaman jd eror sistem sedang sibuk, aplikasi ini mau nipu mau jual data org lain?

Reviews that have been cleaned undergo a tokenization process, where the text is broken down into its smallest units (tokens) [10]. Stopword removal and stemming techniques are also applied to reduce data complexity and increase feature relevance [11]. Table 2 shows the results of data pre-processing.

Table 2. Result of data pre-processing

Content	Stemed	Sentiment
Pengajuan saya sudsh ditolak 2x dan saya tidak pernah terlambat untuk membayar Apa alasan ada kami mengurangkan kredit score tanpa spengetahuan saya limit saya jadi turun dan disuruh menambahkan informasi untuk meningkatkan limit Maksudnya apa ?? Semau hati mengurangkan credit score saya 8x saya mengajukan 8x kali saya tidak pernah telat membayar Kebijakan apa yg terjadi dengan pihak ada kami ??? Kalau tidak bisa membantu lg mohon data saya jangan disebar luaskan TRIMAKASIH	[‘aju’, ‘sudsh’, ‘tolak’, ‘2x’, ‘pernah’, ‘lambat’, ‘bayar’, ‘apa’, ‘alas’, ‘ada’, ‘kami’, ‘kurang’, ‘kredit’, ‘score’, ‘tanpa’, ‘spengetahuan’, ‘limit’, ‘jadi’, ‘turun’, ‘suruh’, ‘tambah’, ‘informasi’, ‘tingkat’, ‘limit’, ‘maksud’, ‘apa’, ‘mau’, ‘hati’, ‘kurang’, ‘credit’, ‘score’, ‘8x’, ‘aju’, ‘8x’, ‘kali’, ‘pernah’, ‘telat’, ‘bayar’, ‘bijak’, ‘apa’, ‘yg’, ‘jadi’, ‘pihak’, ‘ada’, ‘kami’, ‘kalau’, ‘batu’, ‘lg’, ‘mohon’, ‘data’, ‘jangan’, ‘sebar’, ‘luas’, ‘trimakasih’]	Negatif
Aplikasi apaan ni? Dpt rekom kalo aplikasi ini cpat prosesnya, lgsg download malah gagal, trus disuruh perbaikin score credit goblok, baru juga mau minjem, ga jelas, kalo ga ada duit buat minjemin gausa buat aplikasi pinjol. Wkwkwk	[‘aplikasi’, ‘apa’, ‘ni’, ‘dpt’, ‘rekom’, ‘kalo’, ‘aplikasi’, ‘cpat’, ‘proses’, ‘lgsg’, ‘download’, ‘malah’, ‘gagal’, ‘trus’, ‘suruh’, ‘perbaikin’, ‘score’, ‘credit’, ‘goblok’, ‘baru’, ‘mau’, ‘minjem’, ‘ga’, ‘jelas’, ‘kalo’, ‘ga’, ‘ada’, ‘duit’, ‘buat’, ‘minjemin’, ‘gausa’, ‘buat’, ‘aplikasi’, ‘pinjol’, ‘wkwkwwk’]	Negatif
Pinjol bedebah…promo2 versi terbaru ketika install mau ngajukan ulang ..limit anda kurang…kgak usah sms….kgak usah promo ke saya….mirip orang kaya mau bagi2 rejeki niatnya cuma pamer….biar orang pada install.. Pinjol taik …pelanggaran kode etik…tutup aja nih aplikasi…OJK SWI mana suaranya…	[‘pinjol’, ‘bedebahpromo2’, ‘versi’, ‘baru’, ‘ketika’, ‘install’, ‘mau’, ‘ngajukan’, ‘ulang’, ‘limit’, ‘anda’, ‘kurangkgak’, ‘usah’, ‘smskgak’, ‘usah’, ‘promo’, ‘sayamirip’, ‘orang’, ‘kaya’, ‘mau’, ‘bagi2’, ‘rejeki’, ‘niat’, ‘cuma’, ‘pamerbiar’, ‘orang’, ‘install’, ‘pinjol’, ‘taik’, ‘langgar’, ‘kode’, ‘etiktutup’, ‘aja’, ‘nih’, ‘aplikasiojk’, ‘swi’, ‘mana’, ‘suara’]	Negatif
Cukup bagus tapi tolong untk info penagihan bisa melalui notifikasi aplikasi saja atau di SMS, terlbh dahulu sebelum melakukan telepon? Sangat tidak sopan pagi2 telepon dengan nada bicara yang tidak menyenangkan	[‘cukup’, ‘bagus’, ‘tolong’, ‘untk’, ‘info’, ‘tagih’, ‘lalu’, ‘notifikasi’, ‘aplikasi’, ‘saja’, ‘sms’, ‘terlbh’, ‘dahulu’, ‘belum’, ‘laku’, ‘telepon’, ‘sangat’, ‘sopan’, ‘pagi2’, ‘telepon’, ‘nada’, ‘bicara’, ‘senang’]	Negatif
Tolong data saya hapus smuanya sya gamau data sya dipake yg ga bener soalnya aneh pas pertama verifikasi data biasa aja jalan normal ga gangguan,, knpa pas mau mengajukan pinjaman jd eror sistem sedang sibuk, aplikasi ini mau nipu mau jual data org lain?	[‘tolong’, ‘data’, ‘hapus’, ‘smuanya’, ‘sya’, ‘gamau’, ‘data’, ‘sya’, ‘dipake’, ‘yg’, ‘ga’, ‘bener’, ‘soal’, ‘aneh’, ‘pas’, ‘pertama’, ‘verifikasi’, ‘data’, ‘biasa’, ‘aja’, ‘jalan’, ‘normal’, ‘ga’, ‘ganggu’, ‘knpa’, ‘pas’, ‘mau’, ‘aju’, ‘pinjam’, ‘jd’, ‘eror’, ‘sistem’, ‘sedang’, ‘sibuk’, ‘aplikasi’, ‘mau’, ‘nipu’, ‘mau’, ‘jual’, ‘data’, ‘org’, ‘lain’]	Negatif

The first base model, Random Forest is an algorithm that uses several decision trees to make more accurate and stable predictions [12]. Each tree is trained with a random subset of the training data and features, and is not pruned. In classification, the class most frequently predicted by all trees becomes the final prediction [13]. In Classification, the Random Forest output can be expressed as follows:

$$Y = mode\{Y_{1,} Y_{2,} \dots Y_3\} \tag{1}$$

Where:
Y : final prediction
$Y_1, Y_2, \dots Y_3$: predictions from each tree

$$Y = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{2}$$

Where:
Y : average prediction
Y_i : prediction from the *i-th* tree

.

The second base model, Naive Bayes, is a machine learning algorithm based on the application of Bayes' Theorem with the strong assumption that predictors are mutually independent. This algorithm calculates the probability of a particular outcome based on a set of features, taking advantage of the simplicity of the independent feature assumption [14]. This makes the algorithm not only computationally efficient but also highly accurate, especially in scenarios with large data and many features. The mathematical formulation of Naive Bayes is expressed as:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{3}$$

Where:
P(A|B) : posterior probability of class A given predictor B
P(A|B) : likelihood which is the probability of predictor given class
P(A) : prior probability of class A
P(B) : prior probability of predictor B

The third base model, Support Vector Machine (SVM) is a machine learning algorithm that searches for hyperplanes in N-dimensional space to classify data points. SVM can handle linear and non-linear data, and uses kernel functions for data transformation [15]. The data points that influence the position of the hyperplane are called support vectors, with the following formula:

$$min_{w,b} \frac{1}{2} ||w||^2 \tag{4}$$

Subject to the constraints:

$$y_i(w \cdot x_i + b) \geq 1, \forall i \tag{5}$$

where
w : weight vector
b : bias
x_i : training samples
y_i : class labels

Stacking ensemble with meta-models such as Logistic Regression is an approach in machine learning to improve the accuracy of predictive performance [15]. The procedure begins with training basic models of Random Forest, Naive Bayes, and Support Vector Machine (SVM), on the dataset. Each base model, makes predictions independently. The resulting predictions are not used as final results but as input features for the meta-model. The Logistic Regression Meta-model then takes these predictions and analyzes them to understand how each basic model contributes to the overall prediction as shown in Figure 2.

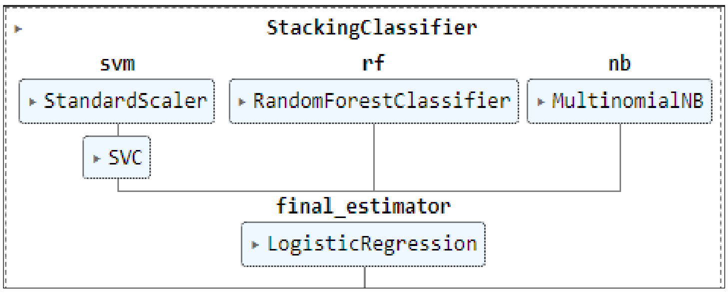

Figure 2. The framework of stacking ensemble classifier

In essence the model learns which ones perform best under any circumstances, thereby synthesizing more accurate predictions.

$$P(Y = 1|x) = \frac{1}{1+e^{-(w^T x+b)}} \tag{6}$$

Where:
$P(Y=1|x)$: probability of the target variable being 1 given the predictors x
w : weight vector
b : bias

Models are evaluated based on metrics such as accuracy, precision, recall, and F1-score. This evaluation was carried out through cross validation to ensure the reliability of the results [16-20]. Results from each base model and stacking ensemble model are compared to determine performance improvements, if any.

The results obtained are analyzed to understand to what extent ensemble stacking improves sentiment classification. Interpretation of the results focuses on how various models handle different aspects of user reviews, and how stacking ensemble helps in obtaining more accurate and robust predictions.

3. RESULTS AND DISCUSSIONS

The confusion matrix presented visualizes the performance of the Random Forest classifier applied to sentiment analysis of online loan application reviews. The matrix indicates that, out of the total samples, 1,184 true positive predictions were made, where the model correctly identified positive sentiments. Conversely, 528 true negatives were identified, where the model accurately predicted negative sentiments. However, the model also produced 229 false negatives, instances where negative sentiments were incorrectly classified as positive, and 59 false positives, where positive sentiments were misclassified as negative. This results in a high number of accurate predictions for positive sentiments, whereas a considerable amount of negative sentiments were misclassified, suggesting a bias in the model towards positive classifications or an imbalance in the dataset. Figure 3 shows the confusion matrix which describe the performance of Random Forest classification model.

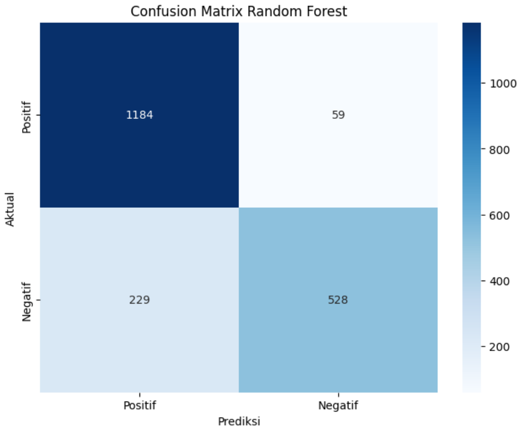

Figure 3. Random Forest confusion matrix

The confusion matrix for the Naive Bayes classifier delineates its performance in sentiment classification of online loan application reviews. It reveals that the classifier correctly predicted 1,196 positive sentiments as true positives and correctly identified 515 negative sentiments as true negatives. The model, however, also incorrectly classified 242 instances as negative that were actually positive (false negatives), and it misclassified 47 positive instances as negative (false positives).

When compared to the Random Forest classifier previously analyzed, the Naive Bayes classifier exhibits a slightly higher number of true positives (1,196 compared to 1,184) and fewer false positives (47 compared to 59), suggesting a marginally better precision in identifying positive sentiments. Conversely, it displays a higher number of false negatives (242 compared to 229), indicating a slightly lower sensitivity or recall for negative sentiments, as shown as confusion matrix in Figure 4.

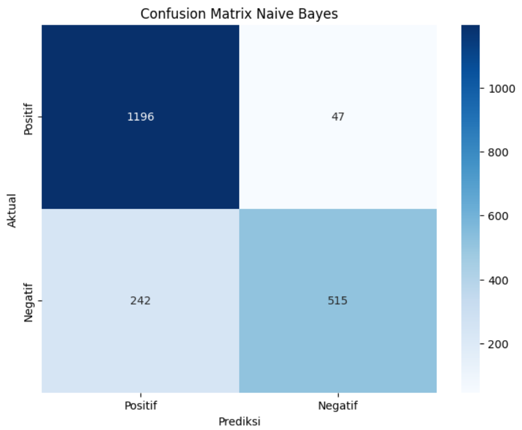

Figure 4. Naïve Bayes confusion matrix

.

The Support Vector Machine (SVM) confusion matrix depicts the classification results for sentiment analysis on online loan application reviews. The matrix indicates that the SVM model correctly classified 1,145 instances as positive sentiments (true positives) and accurately detected 585 instances as negative sentiments (true negatives). However, there were 172 false negatives where negative sentiments were incorrectly classified as positive, and 98 false positives where positive sentiments were mistakenly identified as negative.

Compared to the Random Forest classifier, which had 1,184 true positives and 528 true negatives, the SVM model identified fewer true positives but more true negatives, suggesting better performance in correctly identifying negative sentiments. When contrasted with the Naive Bayes classifier, which produced 1,196 true positives and 515 true negatives, the SVM had fewer true positives but again more true negatives, reinforcing the observation that SVM may be better at classifying negative sentiments than Naive Bayes, as shown as confusion matrix in Figure 5.

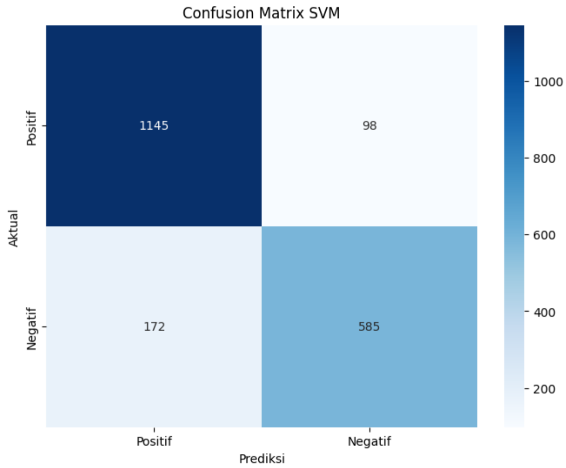

Figure 5. SVM confusion matrix

The confusion matrix for the stacking ensemble model illustrates the classification outcomes for the sentiment analysis of online loan application reviews. The matrix shows that the model has achieved 1,151 true positive classifications for positive sentiment and 590 true negative classifications for negative sentiment. It also incorrectly classified 167 instances as positive that were actually negative (false negatives) and misclassified 92 instances as negative that were truly positive (false positives).

When compared to the performance of the Random Forest, Naive Bayes, and SVM classifiers, the stacking ensemble method offer an improved balance between detecting both positive and negative sentiments. It has a lower false negative rate than the Random Forest and SVM models, and a comparable false positive rate to the Naive Bayes model. Moreover, the true negative rate is superior to that of the Random Forest and Naive Bayes, and on par with the SVM, indicating robustness in identifying negative reviews. This suggests that the stacking ensemble model, with Logistic Regression as the meta-classifier, may provide a more accurate and balanced classification of sentiments in online loan application reviews than the individual models evaluated separately, as shown as confusion matrix in Figure 6.

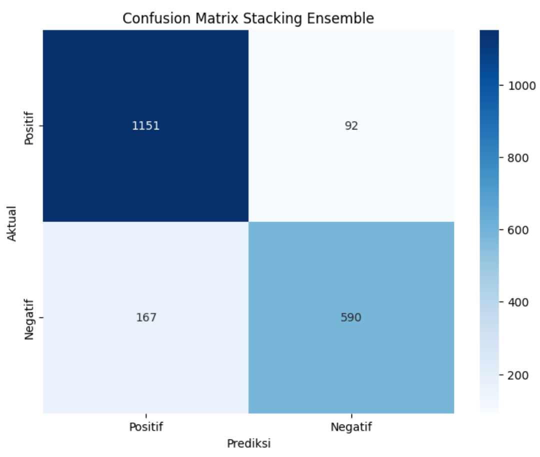

Figure 6. Stacking ensemble confusion matrix

After knowing the values of True Positives, True Negatives, False Positives, False Negatives from the Random Forest, Naive Bayes, SVM and Stacking Ensemble classification models, the evaluation metric values of accuracy, precision, recall and F1-Score can be calculated. Table 1 provides an overview of how each model performs in terms of accuracy, precision, recall, and F1-score for both classes (Negative and Positive).

The Random Forest model demonstrates robust accuracy (85.6%) and precision (83.79%), alongside a commendable recall (90.52%) and F1-score (89.16%). This indicates a strong overall performance with a slight inclination towards correctly predicting positive instances.

The Naive Bayes classifier, while exhibiting a lower accuracy (85.55%) and precision (83.17%) than Random Forest, achieves the highest recall (96.22%) among the models. This suggests a heightened sensitivity in detecting true positives. The F1-score (89.22%) of Naive Bayes is marginally superior to Random Forest, reflecting a balanced precision-recall trade-off despite slightly lower accuracy and precision.

SVM stands out with the highest accuracy (86.5%), suggesting its overall superiority in correctly classifying both positive and negative instances. With precision (86.94%) also being the highest, SVM shows a strong capability to correctly predict positive sentiments while avoiding false positives. However, the recall (90.12%) is slightly lower than Random Forest, indicating a minor trade-off between precision and recall. Nonetheless, the F1-score (89.45%) is very competitive, pointing to its efficacy in harmonic mean between precision and recall.

The stacking ensemble model with Logistic Regression as the meta-classifier presents a close accuracy (87.05%) which is competitive with SVM. It has a precision rate (87.33%) that suggests a strong predictive performance on positive sentiments. The recall (92.6%) is lower than Naive Bayes but higher than both Random Forest and SVM, indicating better sensitivity than the former two. The F1-score (89.89%) is reflective of a balanced precision and recall, suggesting that the ensemble method effectively combines the strengths of the individual models.

Compared to prior research that successfully improved accuracy from 87.72% to 91.22% using a stacking ensemble for depression prediction among university students [9], this study has also successfully enhanced the accuracy of stacking ensemble methods for online loan application reviews, albeit with varying percentage rates as depicted in Table 3. This underscores the versatility and effectiveness of stacking ensemble techniques in different application domains, further substantiating the method's robustness in enhancing predictive performance over singular model approaches.

Table 3. Comparison of classification results

Model	Accuracy	Precision	Recall	F-1 Score
Random Forest	0.856	0.8379	0.9052	0.8916
Naive Bayes	0.8555	0.8317	0.9622	0.8922
SVM	0.865	0.8694	0.9012	0.8945
Stacking Ensemble (LR)	0.8705	0.8733	0.926	0.8989

4. CONCLUSION

In conclusion, this study has embraced the digital transformation in financial services by leveraging user-generated content on platforms like Playstore to gain insights into customer satisfaction and perception of online loan services. Our investigation utilized an advanced stacking ensemble method, integrating the distinct capabilities of Random Forest, Naive Bayes, and SVM classifiers with Logistic Regression as the meta-classifier, to address the limitations and biases inherent in individual machine learning models. This methodological approach has not only enhanced the predictive accuracy of sentiment analysis but also offered a scalable solution to handle the vast volume and complexity of user reviews.

Reflecting on the success of previous research in the field of mental health, which utilized similar ensemble techniques to predict depression with high accuracy, our study confirms the efficacy of these methods in a different domain. By applying a comparable ensemble approach to online loan service reviews, we have furthered the understanding of automated sentiment analysis in the financial sector.

Future research should explore the extension of this model to other domains within the digital service sector, the incorporation of additional machine learning algorithms into the ensemble to explore potential performance gains, and the application of the model to real-time data for dynamic sentiment tracking. Additionally, addressing potential ethical implications related to the privacy of user data and transparency of the analysis will be crucial as the use of automated sentiment analysis continues to expand.

.

REFERENCES

[1] L. Rahmawati and D. B. Santoso, "Implementasi Metode Naive Bayes Untuk Klasifikasi Ulasan Aplikasi E-Commerce Tokopedia," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 6, no. 1, pp. 116–124, Feb. 2023, doi: 10.31539/intecoms.v6i1.5515.

[2] A. Chader, L. Hamdad, and A. Belkhiri, "Sentiment Analysis in Google Play Store: Algerian Reviews Case," 2021, pp. 107–121. doi: 10.1007/978-3-030-58861-8_8.

[3] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Comput Sci Rev*, vol. 41, p. 100413, 2021.

[4] S. N. Singh and T. Sarraf, "Sentiment analysis of a product based on user reviews using random forests algorithm," in *2020 10th International conference on cloud computing, data science & engineering (Confluence)*, 2020, pp. 112–116.

[5] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[6] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A Novel Selective Naive Bayes Algorithm," *Knowl Based Syst*, vol. 192, p. 105361, 2020.

[7] J. Wu, P. Guo, Y. Cheng, H. Zhu, X.-B. Wang, and X. Shao, "Ensemble generalized multiclass support-vector-machine-based health evaluation of complex degradation systems," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2230–2240, 2020.

[8] A. Al Shorman, H. Faris, and I. Aljarah, "Unsupervised intelligent system based on one class support vector machine and Grey Wolf optimization for IoT botnet detection," *J Ambient Intell Humaniz Comput*, vol. 11, pp. 2809–2825, 2020.

[9] A. Daza Vergaray, J. C. H. Miranda, J. B. Cornelio, A. R. López Carranza, and C. F. Ponce Sánchez, "Predicting the depression in university students using stacking ensemble techniques over oversampling method," *Inform Med Unlocked*, vol. 41, p. 101295, 2023, doi: 10.1016/j.imu.2023.101295.

[10] T. Wolf *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[11] Y. A. Alhaj, M. A. A. Al-qaness, A. Dahou, M. Abd Elaziz, D. Zhao, and J. Xiang, "Effects of light stemming on feature extraction and selection for arabic documents classification," *Recent Advances in NLP: The Case of Arabic Language*, pp. 59–79, 2020.

[12] C. N. Noviyanti and A. Alamsyah, "Early Detection of Diabetes Using Random Forest Algorithm," *Journal of Information System Exploration and Research*, vol. 2, no. 1, 2024.

[13] G.-W. Cha *et al.*, "Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets," *Int J Environ Res Public Health*, vol. 17, no. 19, p. 6997, 2020.

[14] S. S. Bafjaish, "Comparative analysis of Naive Bayesian techniques in health-related for classification task," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 1–10, 2020.

[15] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, pp. 341–357, 2020.

[16] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020, pp. 79–91.

[17] M. A. Muslim *et al.*, "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning *," *Intelligent Systems with Applications*, vol. 18, no. December 2022, p. 200204, 2023, doi: 10.1016/j.iswa.2023.200204.

[18] M. A. Muslim, Y. Dasril, H. Javed, W. F. Abror, D. A. A. Pertiwi, and T. Mustaqim, "An Ensemble Stacking Algorithm to Improve Model Accuracy in Bankruptcy Prediction," *Journal of Data Science and Intelligent Systems*, vol. 1, no. 1, 2023.

[19] R. Rofik, R. Aulia, K. Musaadah, S. S. F. Ardyani, and A. A. Hakim, "The Optimization of Credit Scoring Model Using Stacking Ensemble Learning and Oversampling Techniques," *Journal of Information System Exploration and Research*, vol. 2, no. 1, 2024.

[20] A. U. Dullah, F. N. Apsari, and J. Jumanto, "Ensemble learning technique to improve breast cancer classification model," *Journal of Soft Computing Exploration*, vol. 4, no. 2, 2023.