VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

KNOWLEDGE DISCOVERY OF PUBLIC SATISFACTION ON E-HEALTH PLATFORM THROUGH ENSEMBLE LEARNING AND THEMATIC ANALYSIS

Syakillah Nachwa[1]; Ken Ditha Tania[1*]; Naretha Kawadha Pasemah Gumay[1]

Information System[1]
Universitas Sriwijaya, Palembang, Indonesia[1]
www.unsri.ac.id[1]
syakillahnachwa081@gmail.com, kenya.tania@gmail.com*, narethakawadha@unsri.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract—*As Indonesia's national digital health platform, the SATUSEHAT application is central to the country's healthcare integration. However, maintaining user satisfaction remains a challenge. This study evaluates public sentiment by analyzing over 30,000 Google Play Store reviews using an ensemble approach of machine learning and thematic analysis. To address substantial data imbalance, five base classifiers were compared with three ensemble models using Random Under-Sampling and SMOTE. Results indicate that the Stacking Ensemble with SMOTE outperformed all other models, achieving an Accuracy of 91.1% and an AUC-ROC of 0.919. Sentiment analysis reveals a critical dissatisfaction rate, with 83.3% of reviews being negative. By mapping user complaints to the ISO/IEC 25010 software quality framework, it was identified that functional suitability and reliability account for 80.7% of the reported issues. This research contributes a robust methodology for Indonesian-language sentiment analysis and demonstrates the utility of ISO/IEC 25010 in translating raw user feedback into actionable software engineering requirements. Practically, the findings provide the Ministry of Health with an evidence-based roadmap to resolve critical barriers in authentication, OTP delivery, and certificate retrieval.*

Keywords: *Ensemble Learning, ISO/IEC 25010, SATUSEHAT, Sentiment Analysis, Thematic Analysis.*

Intisari—*Sebagai platform kesehatan digital nasional Indonesia, aplikasi SATUSEHAT memegang peran krusial dalam integrasi layanan kesehatan. Namun, besarnya cakupan pengguna menghadirkan tantangan dalam menjaga kepuasan publik. Penelitian ini mengevaluasi sentimen publik dengan menganalisis lebih dari 30.000 ulasan Google Play Store menggunakan pendekatan ensemble machine learning dan analisis tematik. Untuk mengatasi ketidakseimbangan data yang ekstrem, penelitian ini membandingkan lima model tunggal dan tiga model ensemble dengan strategi Random Under-Sampling dan SMOTE. Hasil menunjukkan bahwa Stacking Ensemble dengan SMOTE memberikan performa terbaik dengan Akurasi 91,1% dan AUC-ROC 0,919. Analisis sentimen mengungkapkan tingkat ketidakpuasan yang signifikan, yakni 83,3% ulasan bersifat negatif. Dengan memetakan keluhan tersebut ke dalam kerangka kualitas perangkat lunak ISO/IEC 25010, ditemukan bahwa masalah kelayakan fungsional (functional suitability) dan keandalan (reliability) mendominasi sebesar 80,7%. Melalui standar ISO/IEC 25010, penelitian ini berhasil menyusun metode analisis sentimen yang mampu memetakan keluhan pengguna menjadi perbaikan sistem yang terstruktur. Hasilnya merupakan panduan strategis bagi Kementerian Kesehatan dalam membenahi masalah login, OTP, dan akses sertifikat. Secara praktis, temuan ini memberikan rekomendasi bagi Kementerian Kesehatan untuk memprioritaskan perbaikan pada sistem autentikasi, pengiriman OTP, dan pengunduhan sertifikat.*

Kata Kunci: *Ensemble Learning, ISO/IEC 25010, SATUSEHAT, Analisis Sentimen, Analisis Tematik.*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035

INTRODUCTION

Indonesia's 2024–2029 national agenda (Asta Cita) prioritizes health and technology through digital public service reform, in accordance with Presidential Regulation No. 95/2018 on SPBE [1][2]. The COVID-19 pandemic exposed critical weaknesses in the healthcare system while simultaneously accelerating digital health adoption [3]. In response, the Ministry of Health issued Regulation No. 21/2020, culminating in the SATUSEHAT platform launched on March 1, 2023, replacing Peduli Lindungi and establishing a centralized digital health records system that eliminates the need for physical documents nationwide [4][5][6]. Given its national scope, the platform's success hinges on widespread adoption and service quality, as guaranteed under Article 28H(1) of the 1945 Constitution [7].

Sentiment analysis is a scalable, data-driven approach for evaluating public satisfaction with digital services by interpreting user perceptions from textual reviews [8][9]. With the rapid growth of online reviews, it has become an efficient tool for large-scale public opinion analysis and has been widely applied in assessing digital public services [10][11].

Prior studies demonstrate varying performance across methods. Single-model approaches, such as Naïve Bayes and Support Vector Machine (SVM), have shown strong results, with SVM often outperforming several other single-classifier methods in classification accuracy [12][13]. Meanwhile, ensemble methods such as voting, bagging, stacking, and boosting have also been widely explored in these sectors, including digital banking applications, e-commerce platforms, and public employment systems, where they achieve high classification performance [13][14][15][16][17].

Complementary techniques, such as N-gram [7][18] or NVivo [19] analyses, further reveal recurring user concerns, including system instability and usability issues. However, a significant research gap remains regarding the national health sector. Furthermore, existing studies often treat sentiment classification and issue identification in isolation, rarely aligning findings with formal software quality frameworks.

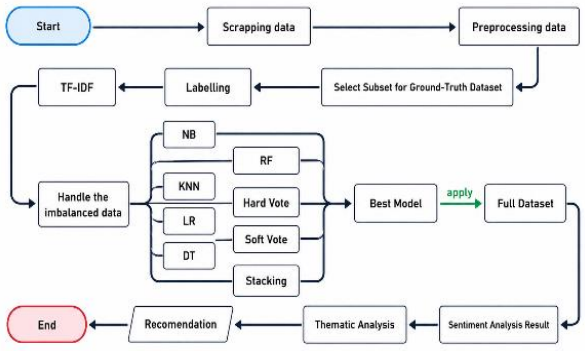
This study addresses those gaps through three contributions:

1. Methodological Comparison: A systematic evaluation of single versus ensemble classifiers combined with class-balancing strategies to identify the optimal model for Indonesian-language reviews.
2. Thematic Integration: The use of N-gram techniques to extract dominant user concerns, bridging the gap between quantitative sentiment and qualitative feedback.
3. Standardized Mapping: The mapping of identified issues to ISO/IEC 25010 quality dimensions to provide structured, actionable recommendations for improving Indonesia's digital health infrastructure.

This study addresses the following research questions: (1) How do single-model and ensemble approaches differ in sentiment classification performance? (2) What is the prevailing sentiment among SATUSEHAT users? (3) Which factors most significantly contribute to public dissatisfaction.

MATERIALS AND METHODS

The methods employed in this research comprisess multiple stages, as illustrated in Figure 1.



Source: (Research Result, 2025)

Figure 1. Research Procedure

Data Collection

A total of 31,363 valid reviews were extracted from the Google Play Store on June 2, 2025, using the google-play-scraper library. The dataset spans from March 1, 2023, to June 2, 2025. To ensure data quality, duplicate and non-Indonesian reviews were removed. Subsequently, text cleaning was performed using regular expression-based Python processing to eliminate usernames, emojis, special characters, and numerical digits. After filtering and cleaning, 31,363 valid user reviews from March 1, 2023, to June 2, 2025, were retained for further analysis.

Data Preprocessing

The preprocessing stage was conducted using Python and standard natural language processing libraries. The procedure consisted of

several key steps. First, case folding was applied to convert all text into lowercase. Next, informal and non-standard words were normalized using an Indonesian slang dictionary. The text was then tokenized into individual words, followed by stemming using the Nazief–Adriani algorithm implemented in the Sastrawi library. Finally, stopword removal was performed using an Indonesian stopword list to eliminate less informative terms. The overall preprocessing workflow is illustrated in Figure 2.

Source: (Research Result, 2025)
Figure 2. The Preprocessing Pipeline

This process produced a clean and standardized textual dataset suitable for further analysis. An example of the preprocessing results at each stage is presented in Table 1.

Table 1. Preprocessing Result

No	Steps	Result
1	Raw	Apps sampaherror terus
2	Case Folding	apps sampaherror terus
3	Normalize	aplikasi sampah error terus
4	Tokenize	[‘aplikasi’, ‘sampah’, ‘error’, ‘terus’]
5	Stemming	['aplikasi', 'sampah', 'error', 'terus']
6	Stopword	aplikasi sampah error

Source: (Research Result, 2025)

Data Annotation

From an initial collection of 31,363 reviews, a preliminary ground-truth dataset containing 21,237 entries (70%) was generated using the pre-trained mdhugol/indonesia-bert sentiment classifier, which categorises Indonesian reviews into positive or negative classes, consisting of 18,092 negative and 3,704 positive reviews.

To ensure reliability and reduce potential bias from automated labeling, a random subset of 400 reviews was manually evaluated by three validators, representing a 5% margin of error. Each validator assessed whether the predicted sentiment label aligned with the contextual meaning of the review. The evaluation yielded an average accuracy of 92%. Inter-annotator agreement, measured using Cohen’s κ, reached 0.9146, indicating almost perfect agreement according to established reliability standards.

Data Vectorization

TF-IDF vectorization was applied with unigrams and bigrams (ngram_range=(1,2)), while tokens appearing in fewer than 5 documents or more than 80% of documents were excluded (min_df=5, max_df=0.8).

Training Models

To mitigate bias and address class imbalance in sentiment data, resampling techniques were applied to the training dataset. SMOTE was chosen for its ability to synthetically generate new minority-class samples to address underrepresentation, while Random Under Sampling was used to reduce the number of majority-class samples to prevent dominance. These choices align with previous research indicating their effectiveness in handling imbalanced data in sentiment analysis [20]. The distribution of the dataset before and after applying the balancing techniques is presented in Table 2.

Table 2. The Balancing Result

Methods	Label	Original Data	Balanced Data
Random Under Sampling	Negative	18092	3704
	Positive	3704	3704
SMOTE	Negative	18092	18092
	Positive	3704	18092

Source: (Research Result, 2025)

This study trained both single and ensemble models for sentiment classification. Single models included Naive Bayes, K-Nearest Neighbours, Logistic Regression, Decision Tree, and Random Forest using 10-fold cross-validation to ensure robustness [21]. Multinomial Naïve Bayes was selected due to its demonstrated effectiveness in prior research [22]. For Ensemble Learning, three approaches were implemented. Hard voting determined the final class using the majority voting among base classifiers, while soft voting averaged the predicted probabilities to select the most likely class [14]. Stacking combined predictions from NB, KNN, DT, and RF as the base models, with LR as the meta-model for final prediction.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035

Sentiment Analysis

After all models were constructed, their performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The F1-score was prioritized as it represents harmonized performance across precision and recall. Each metric was reported as the mean across 10-fold cross-validation to indicate the stability of performance. The model with the highest mean F1-score was selected as the best-performing model and was then used to classify 31.363 user reviews.

Thematic Analysis

Thematic analysis was conducted to understand why users are satisfied or dissatisfied with the app. Using N-gram methods, specifically trigram analysis, recurring three-word phrases were identified to represent complex user issues [23]. The extracted trigrams were separated into positive and negative groups based on sentiment labels. To further interpret user complaints, only the negative trigrams were examined in depth by clustering semantically similar phrases, resulting in 10 themes that comprehensively represent the main user problems.

Table 3 lists each theme with its representative keywords. These themes correspond to ISO/IEC 25010 quality model categories, as only aspects in user feedback were included. Finally, the categories incorporate those seen in negative trigrams, such as functional suitability, performance efficiency, compatibility, usability, reliability, maintainability, and portability [24].

Table 3. Mapping of Negative Themes to ISO/IEC 25010 Categories

ISO/IEC 25010	Negative Theme	Words in Text
Functional Suitability & Reliability	Unstable App	'error', 'bug', 'crash', 'hang', 'login gagal', 'otp, 'download', 'uninstall'
	Missing Content	'kosong', 'hilang'
Performance Efficiency	Resource Hungry	'berat', 'ram', 'baterai', 'cpu', 'memori penuh'
	Underperformance	'lambat', 'berat', 'slow', 'loading', 'timeout'
Usability	Visually Unappealing	'buram', 'kusam', 'ui', 'ux', 'tampilan'
	Poor Navigation	'menu hilang', 'navigasi', 'pindah halaman', 'tanggal lahir'
	Limited User Control	'jenis kelamin', 'ubah nomor', 'sms', 'whatsapp'
Compatibility & Portability	Cross-Platform Issues	'android', 'kompatibel', 'hp lama', 'hp baru', 'versi'
Maintainability	Inexperienced Developer	'baru belajar', 'newbie', 'amatir', 'developer'

Source: (Research Result, 2025)

As shown in Table 3, most negative themes can be mapped to the ISO/IEC 25010 quality characteristics. One theme, frustrating service experience, does not fall under the ISO/IEC 25010 categories, as it concerns helpdesk responsiveness rather than software quality. However, it is still relevant and is discussed in the recommendations section.

RESULTS AND DISCUSSION

Model Performance

Table 4 shows that the Stacking ensemble with SMOTE achieved the best overall performance (Acc = 0.911, F1 = 0.838, AUC-ROC = 0.919). It outperformed all single classifiers and other ensemble configurations. This is consistent with established findings. Stacking generalizes more effectively by training a meta-learner to combine predictions from diverse base classifiers. It exploits the complementary strengths of base classifiers rather than relying on a single decision boundary [25].

KNN consistently yielded the lowest performance under both strategies (US: F1 = 0.551; SMOTE: F1 = 0.520), which aligns with previous studies reporting that KNN performs poorly in sparse, high-dimensional TF-IDF spaces, where the Euclidean distance becomes ineffective due to the curse of dimensionality [26].

Table 4. Performance of Models

Models	Acc	Prec	Rec	F1-score	AUC-ROC
Random Under Sampling					
NB	0.869	0.767	0.817	0.787	0.914
KNN	0.605	0.596	0.667	0.551	0.680
LR	0.871	0.772	0.848	0.800	0.921
DT	0.785	0.687	0.786	0.705	0.792
RF	0.873	0.774	0.846	0.801	0.918
HV	0.832	0.732	0.832	0.759	0.909
SV	0.832	0.732	0.832	0.759	0.909
Stacking	0.872	0.775	0.856	0.804	0.929
SMOTE					
NB	0.885	0.792	0.818	0.804	0.919
KNN	0.561	0.588	0.656	0.520	0.667
LR	0.897	0.810	0.854	0.829	0.925
DT	0.854	0.744	0.791	0.763	0.799
RF	0.906	0.834	0.833	0.833	0.916
HV	0.888	0.796	0.843	0.816	0.911
SV	0.887	0.796	0.843	0.816	0.911
Stacking	**0.911**	**0.848**	**0.830**	**0.838**	**0.919**

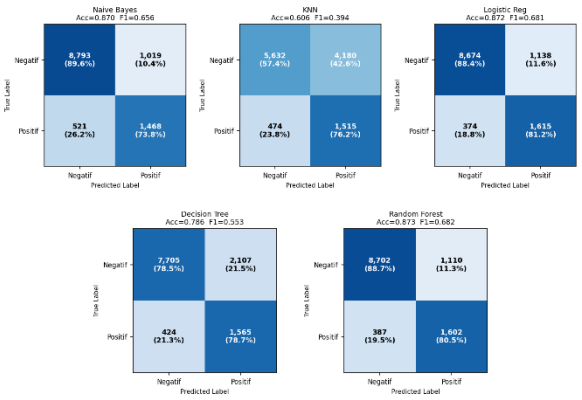
Source: (Research Result, 2025)

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Regarding sampling strategies, SMOTE generally outperformed RUS across most models, particularly for Naive Bayes (F1 =0.787 vs 0.804), Random Forest (F1 = 0.801 vs 0.833) and Logistic Regression (F1 = 0.800 vs 0.829). By generating synthetic minority-class samples instead of removing majority-class data, SMOTE preserved more decision-boundary information and improved the models' ability to classify minority instances effectively [27]. This finding aligns with previous studies showing that SMOTE can improve minority-class sensitivity without substantially reducing overall performance [28]. From a practical perspective, these findings suggest that combining SMOTE with ensemble methods, particularly stacking, can be an effective strategy for real-world sentiment classification tasks involving imbalanced user review data.

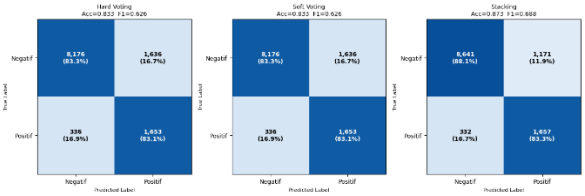
Further analysis using confusion matrices aggregated from all 10 cross-validation folds showed that the single ML models under the RUS framework, Logistic Regression achieved the top performance (Acc = 0.871), yielding an 88.4% True Negative Rate. Naive Bayes and Random Forest performed comparably (Acc ≈ 0.869–0.873), yet both suffered from high False Negative rates, misclassifying 26.2% and 19.5% of positive reviews, respectively. Decision Tree also underperformed with a 21.5% False Positive Rate. The poorest performer was KNN, which misclassified 42.6% of Negative reviews due to the curse of dimensionality in the high-dimensional TF-IDF space.



Source: (Research Result, 2025)

Figure 3. Single Model–RUS Confusion Matrix

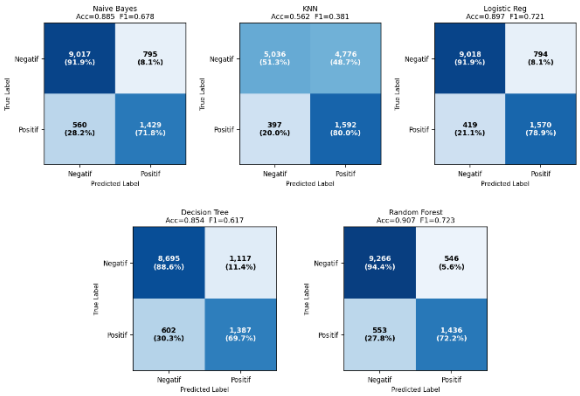
Moving to ensemble methods under RUS, Hard and Soft Voting yielded identical metrics (Acc = 0.832), underperforming compared to the single Logistic Regression model as their False Positive Rate rose to 16.7%. In contrast, Stacking emerged as the superior ensemble variant (Acc = 0.872), securing an 88.1% True Negative Rate and suppressing False Positives better than Voting (11.9% vs. 16.7%). This indicates that the meta-learner in Stacking corrects base classifier weaknesses more effectively than single models.



Source: (Research Result, 2025)

Figure 4. Ensemble Model–RUS Confusion Matrix

A shift to SMOTE oversampling yielded consistent performance gains across almost all single models. Logistic Regression (Acc = 0.897) and Random Forest (Acc = 0.906) stood out as top performers. Notably, Random Forest reduced the False Positive Rate to just 5.6%, the lowest among single models, proving that bagging-based ensembles benefit immensely from synthetic samples. Conversely, KNN remained a severe structural outlier (Acc = 0.562), misclassifying 48.7% of Negative reviews and confirming that its limitations stem from the TF-IDF representation rather than class imbalance.



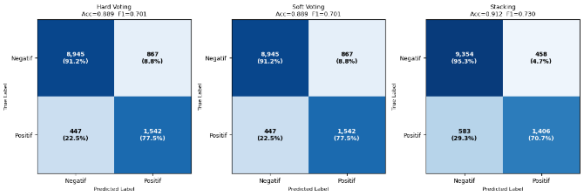
Source: (Research Result, 2025)

Figure 5. Single Model–SMOTE Confusion Matrix

Ultimately, combining ensembles with SMOTE generated the highest overall performance. While Hard and Soft Voting produced identical results (Acc = 0.887 – 0.888), Stacking + SMOTE achieved the absolute best metrics (Acc = 0.911). This configuration recorded the highest True Negative Rate at 95.3%, misclassifying only 4.7% of Negative reviews. Although it exhibited a higher False Negative rate for the Positive class (29.3%) than Voting (22.5%), this trade-off is highly acceptable, as the model successfully prioritizes precision on the majority class (83.3% of the dataset) to ensure reliable sentiment insights.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035



Source: (Research Result, 2025)

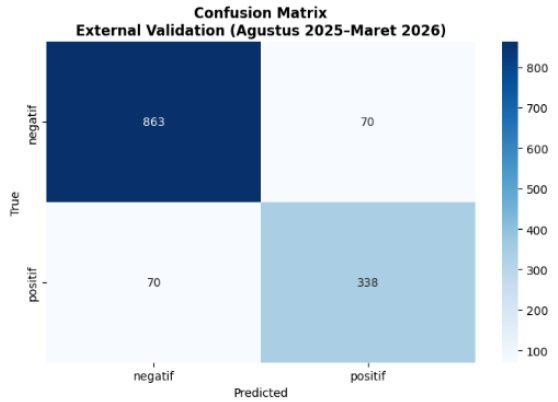
Figure 6. Ensemble Model–SMOTE Confusion Matrix

To verify whether these impressive cross-validation results hold up in practice, the model was tested through temporal external validation using 1,341 unseen SATUSEHAT reviews collected between August 2025 and March 2026. As detailed in Table 5, the model successfully maintained its strong performance (Accuracy = 0.90), effectively confirming its robustness when encountering unseen data from a later time period. This indicates that the proposed model is not only accurate but also stable over time, a critical property for deployment in dynamic real-world environments. The confusion matrix of the temporal external validation is presented in Figure 7.

Table 5. Model Performance under Temporal

	Precision	Recall	F1-Score	Support
Negative	0.92	0.92	0.92	933
Positive	0.83	0.83	0.83	408
Accuracy	-	-	0.90	1341
Macro Avg	0.88	0.88	0.88	1341
Weighted Avg	0.90	0.90	0.90	1341

Source: (Research Result, 2025)



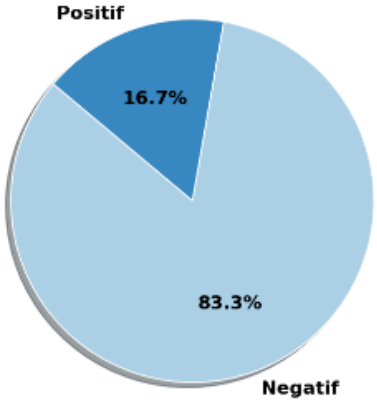
Source: (Research Result, 2025)

Figure 7. Confusion Matrix of External Validation

Sentiment Prediction Result

The Stacking with the SMOTE model showed the highest classification performance. It was then applied to the full dataset of 31,363 user reviews. Results showed that 26,140 reviews were negative and 5,223 were positive. Negative sentiment made up 83.3% of all reviews.

This overwhelming dominance of negative sentiment suggests significant user dissatisfaction. These insights can help policymakers and developers prioritize system improvements, particularly in areas that directly affect user experience and trust in digital health services. The distribution is shown in Figure 8.

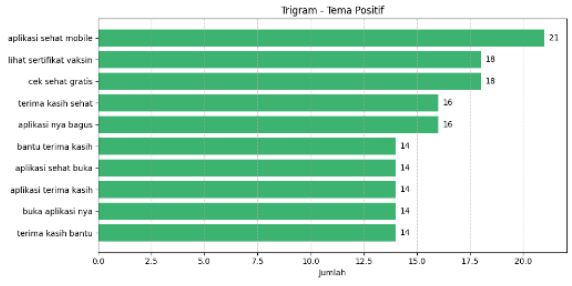


Source: (Research Result, 2025)

Figure 8. Sentiment Prediction Result

Trigram Results

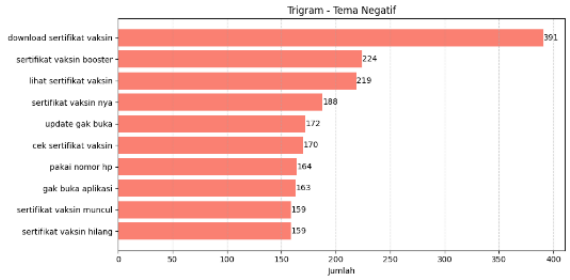
Thematic analysis revealed key insights into user sentiment. Although some positive expressions (Figure 9), such as “app is good” (16 occurrences) and “thank you help” (14 occurrences) were identified, suggesting that the SATUSEHAT application provided value for specific users, but the overall sentiment was predominantly negative.



Source: (Research Result, 2025)

Figure 9. Positive Trigram Results

As illustrated in Figure 10, the most frequently mentioned concerns relate to difficulties in accessing vaccine certificates, with phrases such as “unable to download vaccine certificate” and “unable to view vaccine certificate” appearing hundreds of times. Login failures using phone numbers were also commonly reported. These patterns reflect significant usability and accessibility challenges that users face.



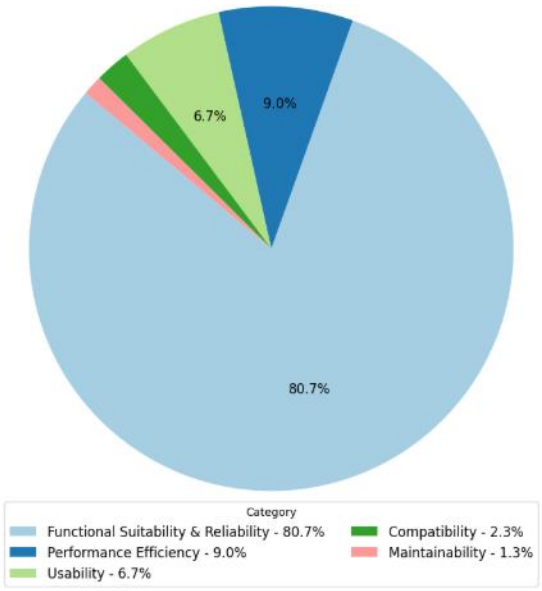
Source: (Research Result, 2025)

Figure 10. Negative Trigram Results

Factors Leading to Negative Reviews

Further analysis of negative trigrams began by filtering out generic or irrelevant examples, such as filler expressions lacking meaningful user feedback. Remaining trigrams were grouped by semantic similarity, consolidating those with similar meanings.

As a result of this process, 10 key themes were identified: unstable app, underperformance, visually unappealing, missing content, cross platform issues, resource hungry, poor navigation, inexperienced developer, frustrating service experience, and limited user control. These themes capture the most common user complaints. The negative themes were then mapped to ISO/IEC 25010 quality characteristics (Figure 11) to provide a structured evaluation of the SATUSEHAT app's weaknesses.



Source: (Research Result, 2025)

Figure 11. Negative Categorization Percentage

1. Functional Suitability and Reliability
   Functional Suitability and Reliability was the most dominant category, accounting for 80.7% of user complaints. Users initially encountered unstable performance and missing content, which disrupted key functions. These issues often caused failures in accessing the app. Notably, 163 users reported "*ga buka aplikasi*" (cannot open), indicating repeated access issues. Many lost personal data after updates. Persistent login troubles, such as not receiving OTP codes, and blocked account access. As a result, many were unable to download vaccination certificates or access the app at all. For example, one user wrote: *"...pas mau verifikasi data gak bisa terus keluar dulu dari aplikasi pas mau log in lagi malah muter muter gak bisa login gimana orang mau nikmati kado ulang tahun pemeriksaan kesehatan gratis...."*.

2. Performance efficiency
   Performance efficiency includes responsiveness, resource use, and stability. After updates, many users reported slower performance. They faced recurring lag, errors, and long loading times. These problems blocked access to key features, such as vaccination certificates.
   Building on these concerns, a total of 574 negative trigrams cited the app's heavy resource use and slow performance. Users found the app cumbersome, reporting slower speeds, reduced reliability, rapid battery drain, and high storage usage, even for basic tasks such as viewing vaccine data. Feedback included statements such as "*Loading lama, sering error*..." and "*Aplikasinya lemot*."

3. Usability
   User dissatisfaction with the redesign centers on confusing navigation and an unattractive visual design. Data show 501 trigram mentions of flawed navigation and 2,029 mentions of design issues, pointing to widespread frustration with both how users navigate the interface and its overall appearance. Representative feedback included: "*Agak membingungkan tampilannya*," "*Butuh penyesuaian lagi dengan tampilan baru, sebenarnya yg lama lebih ok sih, kenapa di ganti yah...*" and "*Pas pilih jenis kelamin gk bisa di klik....*"
   Navigation issues were especially apparent among users with limited digital skills, particularly when performing key tasks like downloading vaccination certificates. Entering a date of birth required excessive scrolling instead of direct input, and users also highlighted limited customization options, such as the absence of dark mode and restricted OTP delivery methods.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035

4. Compatibility and Portability
   Compatibility and Portability refer to the app's ability to perform reliably across diverse devices, operating systems, and technical settings. A total of 1,067 trigram occurrences were associated with cross-platform issues. For instance, while the app operated smoothly on certain models, it failed to launch on others. Additionally, users observed that specific features functioned only with the latest Android releases. One user noted: "*Tolong kebijakan nya tim pengembang aplikasi untuk bisa membuat ketersediaan download untuk HP android versi 5 kebawah, karena tidak semua masyarakat punya HP Android versi 6 keatas*."

5. Maintainability
   Maintainability refers to the ease of updating and maintaining a product over time. It also covers the experience of those managing its upkeep. Many users reported recurring problems with update management and software stability. These issues raised concerns about developer competence and the system's sustainability. One user said, "*Jika aplikasi belum siap digunakan sebaiknya kembalikan dulu ke Peduli Lindungi dan jangan dipaksa untuk dilaunching. Aplikasi mau log in lama banget, bisa beberapa menit akhirnya error lagi. Mohon untuk menjadi perhatian tim developer dan Kemenkes, niatnya udah bagus tapi aplikasi belum mendukung*." Unresolved bugs and poorly tested updates eroded user confidence in the product's reliability and maintenance.

Recommendations for Improvement

A thematic analysis of SATUSEHAT reviews reveals strong public support for digital health services, particularly in terms of accessibility. However, several technical issues were identified. First, regular usability testing is recommended before updates to enhance user interaction and interface design, with prior studies showing improvements of up to 100% in government apps [29]. Second, improving server capacity and ensuring routine updates are essential to minimise crashes and loading delays. Monitoring user reviews is also advised to detect bugs early and sustain public satisfaction.

Beyond technical factors, the reviews also highlighted consistent concerns regarding customer service responsiveness. Although these issues lie outside the ISO/IEC 25010 quality dimensions, they remain critical for maintaining user trust. Approximately 1,736 negative trigrams point to frustration over slow or unanswered helpdesk support. To address this, SATUSEHAT should enhance its support system by providing faster, more empathetic, and user-centered assistance. Clear response-time SLAs, a combined chatbot–human support model, and transparent escalation procedures can help improve public perception and satisfaction. Prior studies also indicate that chatbot systems with human-like communication styles can strengthen user trust and increase satisfaction [30].

CONCLUSION

This study investigated sentiment toward the SATUSEHAT app using machine learning models. The Stacking Ensemble with SMOTE delivered the best performance, achieving an accuracy = 0.911, F1 = 0.838, and AUC-ROC = 0.919, outperforming all single classifiers and other ensemble configurations across both sampling strategies. Confusion matrix analysis further confirmed this superiority, with the Stacking + SMOTE model achieving the highest True Negative Rate of 95.3%, correctly identifying the vast majority of negative reviews that dominate the dataset. This model was applied to 31,363 Google Play Store reviews, revealing that 83.3% of users expressed negative sentiment, demonstrating ongoing public dissatisfaction with the platform.

Sentiment analysis reveals a high level of dissatisfaction, with 83.3% of reviews classified as negative. By mapping user feedback to the ISO/IEC 25010 framework, this study facilitates targeted identification of software quality issues, enabling policymakers and developers to prioritize improvements more effectively. Thematic findings indicate that most issues relate to functional suitability and reliability (80.7%), followed by performance efficiency (9%) and usability (6.7%), with key problems in login, OTP delivery, and certificate retrieval affecting user trust and system usability.

Based on the analysis, several improvements are recommended, including continuous usability evaluation, prioritization of critical technical issues such as login and OTP failures, and enhancement of system scalability. In addition, strengthening user support services is essential to address concerns beyond technical quality and improve overall satisfaction.

Despite these contributions, this study has several limitations. The dataset included only Google Play Store reviews, so it may not fully reflect users' perspectives on platforms like Twitter/X or the App Store. The current approach does not explicitly capture emoji-based sentiment or detect sarcasm or context-dependent language. These

factors may affect interpretation of some user reviews.

To address these limitations, future research should use advanced natural language processing techniques to better capture nuanced semantic meanings in user feedback through contextual word embeddings and Transformer-based models, uncover latent concerns using topic modeling methods such as BERTopic instead of relying solely on N-gram analysis, clarify how sentiment signals are expressed through emojis, develop methods to detect sarcasm and contextual irony in user reviews, and expand the analysis across multiple platforms and over longer time periods to better understand the evolution of public perceptions of digital health services.

REFERENCE

[1] Kementerian Kesehatan Republik Indonesia, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/33/2025 tentang Petunjuk Teknis Pemeriksaan Kesehatan Gratis Hari Ulang Tahun," Jakarta, 2025.

[2] Presiden Republik Indonesia, "Peraturan Presiden Republik Indonesia Nomor 95 Tahun 2018 tentang Sistem Pemerintahan Berbasis Elektronik," Jakarta, 2018.

[3] R. Filip, R. Gheorghita Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: A Review of Pandemic Measures and Problems," *J. Pers. Med.*, vol. 12, no. 8, p. 1295, Aug. 2022, doi: 10.3390/jpm12081295.

[4] Kementerian Kesehatan Republik Indonesia, *Cetak Biru Strategi Transformasi Digital Kesehatan*, vol. 1. Jakarta, 2021.

[5] Kementerian Kesehatan Republik Indonesia, "Peraturan Menteri Kesehatan Nomor 21 Tahun 2020 tentang Rencana Strategis Kementerian Kesehatan Tahun 2020-2024," Indonesia, 2020.

[6] Direktorat Promosi Kesehatan dan Pemberdayaan Masyarakat Kemenkes RI, "PeduliLindungi Resmi Berubah Menjadi SATUSEHAT," Promkes.kemkes.go.id. Accessed: Jun. 23, 2025. [Online]. Available: https://promkes.kemkes.go.id/pedulilindungi-resmi-berubah-menjadi-satusehat

[7] H. Melani Puspasari, I. Zharif Mustaqim, A. Tri Utami, R. Syalevi, and Y. Ruldeviyani, "Evaluation of Indonesia's police public service platforms through sentiment and thematic analysis," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, p. 1596, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1596-1607.

[8] N. Rizun, A. Revina, and N. Edelmann, "Application of Text Analytics in Public Service Co-Creation: Literature Review and Research Framework," in *Proceedings of the 24th Annual International Conference on Digital Government Research*, New York, NY, USA: ACM, Jul. 2023, pp. 12–22. doi: 10.1145/3598469.3598471.

[9] H. D. Sharma and P. Goyal, "An Analysis of Sentiment: Methods, Applications, and Challenges," in *RAiSE-2023*, Basel Switzerland: MDPI, Dec. 2023, p. 68. doi: 10.3390/engproc2023059068.

[10] V. Novalia, K. Ditha Tania, A. Meiriza, and A. Wedhasmara, "Knowledge Discovery of Application Review Using Word Embedding's Comparison with CNN-LSTM Model on Sentiment Analysis," in *2024 International Conference on Electrical Engineering and Computer Science (ICECOS)*, IEEE, Sep. 2024, pp. 234–238. doi: 10.1109/ICECOS63900.2024.10791113.

[11] D. Khoiriyah Harahap, K. Ditha Tania, and P. Eka Sevtiyuni, "Topic Mining-Based Knowledge Discovery of User Health Information Needs," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 4, pp. 641–653, Oct. 2025, doi: 10.35882/ijeeemi.v7i4.270.

[12] D. A. Wulandari, F. A. Bachtiar, and I. Indriati, "Aspect Based Sentiment Analysis on Shopee Application Reviews Using Support Vector Machine," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 15, no. 02, p. 99, Jan. 2025, doi: 10.24843/LKJITI.2024.v15.i02.p03.

[13] Raksaka Indra Alhaqq, I Made Kurniawan Putra, and Yova Ruldeviyani, "Analisis Sentimen terhadap Penggunaan Aplikasi MySAPK BKN di Google Play Store," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 105–113, May 2022, doi: 10.22146/jnteti.v11i2.3528.

[14] B. Andrian, T. Simanungkalit, I. Budi, and A. F. Wicaksono, "Sentiment Analysis on Customer Satisfaction of Digital Banking in Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, 2022, doi: 10.14569/IJACSA.2022.0130356.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7035

[15] E. Junianto, M. Puspitasari, S. I. Zakaria, T. Arifin, and I. W. P. Agung, "Klasifikasi Emosi pada Teks Berbahasa Inggris Menggunakan Pendekatan Ensemble Bagging," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 13, no. 4, pp. 272–281, Nov. 2024, doi: 10.22146/jnteti.v13i4.14440.

[16] H. Herianto, "Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1272–1285, Sep. 2024, doi: 10.47738/jads.v5i3.316.

[17] T. A. Putra, V. Ariandi, and S. Defit, "Enhancing Accuracy by Using Boosting and Stacking Techniques on the Random Forest Algorithm on Data from Social Media X," *ILKOM Jurnal Ilmiah*, vol. 16, no. 2, pp. 184–189, Aug. 2024, doi: 10.33096/ilkom.v16i2.2058.184-189.

[18] I. Zharif Mustaqim, H. Melani Puspasari, A. Tri Utami, R. Syalevi, and Y. Ruldeviyani, "Assessing public satisfaction of public service application using supervised machine learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, p. 1608, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1608-1618.

[19] P. Lee *et al.*, "Digital Health COVID-19 Impact Assessment: Lessons Learned and Compelling Needs," *NAM Perspectives*, Jan. 2022, doi: 10.31478/202201c.

[20] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences*, vol. 13, no. 6, p. 4006, Mar. 2023, doi: 10.3390/app13064006.

[21] V. Lumumba, D. Kiprotich, M. Mpaine, N. Makena, and M. Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," *American Journal of Theoretical and Applied Statistics*, vol. 13, no. 5, pp. 127–137, Oct. 2024, doi: 10.11648/j.ajtas.20241305.13.

[22] C. Dewi, R.-C. Chen, H. J. Christanto, and F. Cauteruccio, "Multinomial Naïve Bayes Classifier for Sentiment Analysis of Internet Movie Database," *Vietnam Journal of Computer Science*, vol. 10, no. 04, pp. 485–498, Nov. 2023, doi: 10.1142/S2196888823500100.

[23] S. K. Ahmed *et al.*, "Using thematic analysis in qualitative research," *Journal of Medicine, Surgery, and Public Health*, vol. 6, p. 100198, Aug. 2025, doi: 10.1016/j.glmedi.2025.100198.

[24] B. I. Rumabar and E. Maria, "Evaluasi Kualitas Shopeepay Menggunakan ISO/IEC 25010," *Jurnal Sistem Informasi Bisnis*, vol. 14, no. 1, pp. 54–61, Jan. 2024, doi: 10.21456/vol14iss1pp54-61.

[25] J. Kazmaier and J. H. van Vuuren, "The power of ensemble learning in sentiment analysis," *Expert Syst. Appl.*, vol. 187, p. 115819, Jan. 2022, doi: 10.1016/j.eswa.2021.115819.

[26] K. Madatov, S. Sattarova, and J. Vičič, "TF-IDF-Based Classification of Uzbek Educational Texts," *Applied Sciences*, vol. 15, no. 19, p. 10808, Oct. 2025, doi: 10.3390/app151910808.

[27] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, no. 1, p. 21631, Jul. 2025, doi: 10.1038/s41598-025-05791-7.

[28] N. Nasaruddin, N. Masseran, W. M. R. Idris, and A. Z. Ul-Saufie, "A SMOTE PCA HDBSCAN approach for enhancing water quality classification in imbalanced datasets," *Sci. Rep.*, vol. 15, no. 1, p. 13059, Apr. 2025, doi: 10.1038/s41598-025-97248-0.

[29] Hafidz. Firdaus and Azizah. Zakiah, "Implementation of Usability Testing Methods to Measure the Usability Aspect of Management Information System Mobile Application (Case Study Sukamiskin Correctional Institution)," *International Journal of Modern Education and Computer Science*, vol. 13, no. 5, pp. 58–67, Oct. 2021, doi: 10.5815/ijmecs.2021.05.06.

[30] Y. Xu, J. Zhang, R. Chi, and G. Deng, "Enhancing customer satisfaction with chatbots: the influence of anthropomorphic communication styles and anthropomorphised roles," *Nankai Business Review International*, vol. 14, no. 2, pp. 249–271, Jun. 2023, doi: 10.1108/NBRI-06-2021-0041.