International Journal of Technology

http://ijtech.eng.ui.ac.id

Research Article

Automated Forensic Diagnostics of Mobile Banking Reliability: A Comparative Study of Engineering Failure Modes in Indonesia and Thailand

Arya Anggadipa Manova[1], Muhardi Saputra[1*], Riska Yanu Fa'rifah[1], Sitthidej Bamrungsap[2]

[1]Department of Information System, School of Industrial Engineering, Telkom University, Bandung 40257, Indonesia

[2]Institute of Management Science, Faculty of Management, Kasetsart University, Bangkok 10900, Thailand

*Corresponding author: muhardi@telkomuniversity.ac.id; Tel.: +6285274258097; Fax.: +62227565200

Abstract: Mobile banking's rapid expansion in Southeast Asia has increased user dependency, making service disruptions a critical source of technostress. However, comparative evidence on the modes of engineering failure across different national ecosystems remains limited. This study aimed to investigate and contrast software reliability issues by analyzing user feedback for two market-leading applications: Livin' by Mandiri (Indonesia) and K PLUS (Thailand). A dataset of 200,000 reviews of Google Play Store was processed using an automated diagnostic framework. This methodology employed Latent Dirichlet Allocation (LDA) to extract latent technical dimensions. In addition, weak supervision via VADER polarity scoring was used to generate pseudo-labels, training the Support Vector Machine (SVM) classifier on the TF-IDF feature to effectively separate failure-related signals from satisfaction reports. The SVM model achieved a strong weighted F1-score of 92.59% and 94.08% for the Indonesian and Thai datasets, respectively. The aspect-based sentiment distribution reveals distinct failure profiles: post-update access barriers and authentication difficulties mostly drive Indonesian user complaints, while client-side runtime instability and application crashes more frequently affect Thai users. Furthermore, the study identifies a security paradox in both ecosystems, where users prioritize seamless service availability over visible data privacy measures. These findings indicate that reliability mitigation strategies should be localized. Indonesian operators should prioritize deployment quality assurance, and Thai operators should optimize client-side resiliency. Both ecosystems require frictionless, backend-driven security measures to minimize user technostress.

Keywords: Latent dirichlet allocation; Mobile banking; Reliability engineering; Support vector machine; Technostress

1. Introduction

Over the past two decades, Southeast Asia's financial landscape has undergone a radical transformation into a highly connected digital ecosystem (Murinde et al., 2022), with mobile banking penetration rates in Indonesia and Thailand exceeding 90% (The Nation Thailand, 2024; Tseng & Koy, 2025; UOB et al., 2024; Visa, 2024). However, this rapidly growing reliance presents a technical paradox: users' tolerance for technical disruptions decreases as they become increasingly dependent on digital platforms. Service disruptions (such as latency, system failures, or authentication errors) can trigger "technostress", a modern adaptation problem arising from individuals' difficulty in balancing new technologies (Bhatt & Kothari, 2022). Financial institutions must shift from traditional, reactive IT operations to proactive Site Reliability Engineering (SRE) to reduce this technostress and handle the massive scale of digital transactions. SRE applies rigorous SRE principles to operations, using automation, continuous monitoring, and predictive analytics to ensure high availability and minimize system downtime (Subramani,

https://doi.org/10.14716/ijtech.v17i4.8422
Received February 2026; Revised April 2026; Accepted June 2026

2025). Mitigating service disruptions is crucial to maintaining customer trust and satisfaction in highly critical environments like cloud infrastructures (Datla, 2023). Therefore, implementing these practices is no longer merely an option but an essential requirement for modern banking architecture to guarantee operational resilience (Subramani, 2025). Ensuring service reliability and mitigating technological friction has become a critical technical challenge for banking institutions to maintain sustainable user adaption (Jadil et al., 2021).

Historically, the evaluation of banking service quality has relied heavily on survey-based instruments. Although theoretically sound, these methods are static, time-consuming, and unable to detect real-time technical anomalies (Adiningtyas & Auliani, 2024), making them insufficient for diagnosing complex system failures within modern agile development ecosystems (Rahman et al., 2022). To address this, recent studies in the field of software engineering emphasize that applying machine learning and big data analysis to user-generated reviews provides a scalable, real-time alternative for evaluating mobile application reliability (Dabrowski et al., 2022). These reviews serve as large-scale participatory sensors that instantly capture user experiences in a rich technical context (Lutfi et al., 2023).

Although the urgency to understand digital service failures is increasingly pressing, the current literature still shows a significant gap in cross-country comparative analysis. Most previous studies have focused only on sentiment analysis within a single country (Adiningtyas & Auliani, 2024; Andrian et al., 2022) or have been limited to analyzing technical security vulnerabilities without considering the user experience dimension (Titiakarawongse et al., 2024). No comprehensive study has compared the engineering failure profiles of Indonesian and Thai banking ecosystems. In fact, differences in the maturity of technological infrastructure and device characteristics in the two countries have the potential to create fundamentally different technostress triggers (Fathimath & Santhi, 2025; Waliszewski & Warchlewska, 2020). The absence of empirical evidence can make it difficult for systems engineers and strategic decision-makers to create precise improvement priorities appropriate to each country's local context. Table 1 summarizes the methodological and contextual gaps in foundational studies compared to the proposed framework to clearly position this research within the existing literature.

Table 1 Comparison of Methodologies and Research Gaps Across Prior Studies and This Research

Study	Methodological Approach	Focus and Context	Identified Gap / Limitation
(Andrian et al., 2022)	9 Standalone Classifiers (e.g., SVM, Naïve Bayes, RF) and 2 Ensemble Methods	Customer satisfaction analysis of three digital banks in Indonesia using Twitter data.	Lacks engineering depth; stops at general sentiment without extracting specific software failure modes or demographic analysis.
(Adiningtyas & Auliani, 2024)	e-ServQual + Naïve Bayes	Measuring e-ServQual dimensions and consumer sentiment across three top mobile banking applications in Indonesia.	Single-country limitation; fails to capture variations in failure patterns across different national technological ecosystems.
(Titiakarawongse et al., 2024)	SAST, OWASP Mobile Top 10	Comparative security vulnerability analysis between Thai and non-Thai mobile banking applications.	Excludes the end-user dimension; detects code vulnerabilities but ignores real-world user impact and technostress triggers.
This Study	LDA, VADER + SVM	Automated cross-country forensic diagnostics of reliability and technostress triggers (Indonesia vs. Thailand) using 200,000 application reviews.	Addresses prior gaps by bridging cross-border analysis with end-user experience to extract actionable reliability engineering priorities.

In response to this gap, this study proposes a comparative, applied analytical framework

to identify technological stress triggers across two distinct banking ecosystems, by utilizing a large dataset of 200,000 user reviews from two market-leading applications: Livin' by Mandiri (Indonesia) and K PLUS (Thailand), which serve tens of millions of active users each (Bank Mandiri, 2025; Kasikorn Bank, 2025). This study challenges the assumption that digital service quality is a monolithic concept. By integrating Latent Dirichlet Allocation (LDA) for topic modelling and Support Vector Machine (SVM) for automatic classification, this study challenges the assumption that digital service quality is monolithic. This study aims to: (1) provide large-scale empirical evidence regarding mobile banking service failures, (2) construct a taxonomy of technostress triggers derived from real-word complaints, and (3) identify reliability engineering priorities tailored to the local context. Ultimately, this study demonstrates how applied data science can bridge the gap between user psychology and reliability engineering, providing a scalable blueprint for mitigating digital service risk. The following research questions guide this study to strengthen the engineering focus and improve the reproducibility of comparative diagnostics:

RQ1: What latent technical dimensions of mobile banking reliability are reflected in large-scale user reviews for Livin' by Mandiri (Indonesia) and K PLUS (Thailand)?

RQ2: How do the dominant failure modes and reliability-related complaints differ between the two national ecosystems?

RQ3: What localized reliability improvement priorities can be derived from cross-country forensic comparison of extracted failure dimensions?

These research questions enable the study to move beyond general sentiment reporting and toward a diagnostic interpretation of user-generated feedback as a proxy signal for software reliability issues in mobile banking platforms.

2. Methods

In this study, we propose an automated diagnostic framework for processing large-scale unstructured feedback in reliability engineering. The methodology follows a systematic pipeline adapted from the Knowledge Discovery in Databases (KDD) process, optimized for high-dimensional textual data. Figure 1 illustrates the overall system architecture comprising five core modules: (1) Data Construction and Preprocessing, (2) Unsupervised Fault Isolation, (3) Pseudo-Labeling and Feature Extraction, (4) Supervised Risk Classification, and (5) Evaluation Metrics.

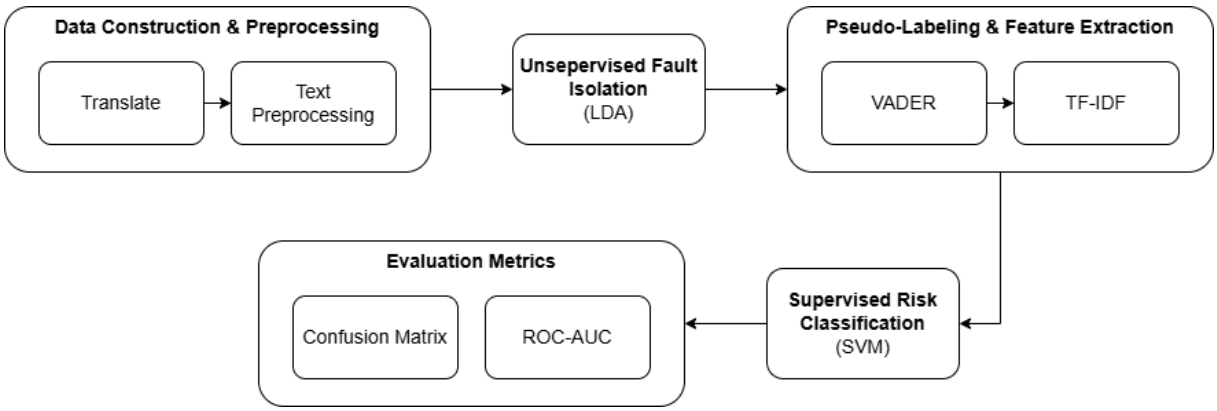


Figure 1 General architecture of the experiment

2.1 Dataset Construction and Preprocessing

This study analyzes a dataset of 200,000 user reviews collected from the Google Play Store, which is equally distributed between two market-leading mobile banking applications: Livin'

by Mandiri (Indonesia) and K PLUS (Thailand). Because the original reviews were written in Indonesian and Thai, all texts were automatically translated into English using the Python *googletrans* library to ensure a uniform linguistic basis for cross-country comparison. Although automatic translation may result in minor semantic noise, recent studies on cross-lingual Natural Language Processing (NLP) have indicated that modern machine translation models are highly effective at preserving core semantic dimensions and sentiment polarity (Alkhushayni & Lee, 2025; Miah et al., 2024). Therefore, this translation step successfully standardized the semantic space for feature extraction without compromising the extracted macro-level technical taxonomy's validity. The automated procedure was consistently applied to both datasets to further reduce bias and maintain methodological symmetry.

Because the dataset consists of publicly available user-generated content, no personally identifiable information was intentionally collected, stored, or processed. The dataset was screened before analysis to remove empty entries, duplicated reviews, and reviews containing no meaningful textual content. A multi-stage preprocessing pipeline was applied after translation to improve text quality and reduce noise. The preprocessing steps included case folding, contraction expansion, non-alphabetic character removal, tokenization, and stopword filtering. Bigram and trigram modeling was performed to capture short technical expressions frequently found in software-related complaints. Finally, lemmatization was applied to normalize inflected words into their base forms, thereby improving feature consistency.

2.2 Automated Fault Isolation Using the LDA Configuration

To implement automatic fault isolation without relying on predefined failure categories, this study employs an unsupervised, generative probabilistic model called Latent Dirichlet Allocation (LDA). In the context of service engineering, LDA functions as a probabilistic topic model that groups co-occurring words into latent topics, which are then used to represent and cluster services or technical artifacts into meaningful dimensions (Geeganage et al., 2024; Yang et al., 2020). The model assumes that each review is a mixture of hidden topics characterized by a specific word distribution (Vayansky & Kumar, 2020). This unsupervised approach is crucial in this study because it allows the *Technostress Trigger Taxonomy* to emerge organically from the data distribution itself, thereby avoiding the bias inherent in manual tagging or rigid survey construction. The implementation uses the *LdaModel* module from the Python *Gensim* library, chosen for its ability to yield consistent, deterministic results during topic generation, ensuring the extracted aspect's reproducibility.

Configuring the LDA model requires a rigorous optimization strategy to ensure that the extracted topics reflect applicable technical insights rather than a random collection of words. The main technical challenge in topic modeling is determining the optimal number of latent topics (K), because setting K too low yields overly broad categories, while setting K too high yields fragmented, uninterpretable clusters. To address this, an iterative hyperparameter tuning process is performed by calculating the Coherence Score (C_v) for values of K ranging from 2 to 10. The C_v Metrics measure the meaningfulness and conceptual usefulness of a topic and its constituent words (Churchill & Singh, 2022). This metric scores a topic by measuring the degree of semantic similarity among the highest-weight words in that topic (Zimmermann et al., 2024).

Additionally, to improve topic quality across heterogeneous review lengths, the Dirichlet hyperparameters α (document–topic density) and η/β (topic–word density) were set to 'auto', allowing the model to estimate asymmetric priors from the corpus. This strategy improves the quality of topics in heterogeneous software engineering texts (e.g., bug reports and user feedback). Therefore, it is well suited to handling variations in review length on the Google Play Store (Silva et al., 2021). By allowing the model to estimate these densities dynamically, the learned prior promotes a sparse topic-word distribution, reducing generic noise and highlighting a limited set of high-probability technical keywords (e.g., *bug*, *timeout*, *crash*), providing useful discriminative features for subsequent error isolation (Watanabe & Baturo, 2024). This opti-

mized LDA configuration successfully separates complex mixtures of user sentiment, isolating specific infrastructure failure points for subsequent risk quantification (Silva et al., 2021).

2.3 Pseudo-Labeling and Feature Extraction

To implement risk measurement in high-capacity engineering workflows, we adopt a weak supervision strategy to generate training labels. Relying on manual annotation of large-scale datasets creates significant operational barriers and introduces subjective variability, hindering diagnostic system reproducibility. To overcome these limitations, the Valence Aware Dictionary and sEntiment Reasoner (VADER) lexicon is applied to generate pseudo-labels. VADER uses a sophisticated rule-based engine that integrates grammatical and syntactic heuristics to handle negation, modifiers, and capitalization, which are commonly found in user-generated content. This algorithm calculates a normalized composite score ranging from -1 (very negative) to +1 (very positive) by summing the valence scores of each word in the text (Arief & Samsudin, 2023). The valence scores are then calculated as follows: A strict threshold protocol is applied to ensure high-confidence training data: reviews with a combined score $\leq$ -0.05 are labeled negative (indicating technostress or service failure). In contrast, those with a score $\geq$ 0.05 are labeled as positive. Reviews that fall between these thresholds are categorized as neutral and excluded from the training set, serving as a noise-reduction filter to sharpen the distinction between functional satisfaction and technical failure. In this framework, pseudo-labeling is treated as a robust operational proxy for separating failure-related signals from satisfaction reports at scale rather than as a definitive psychological truth. Recent research in Natural Language Processing (NLP) indicates that weak lexicon-based supervision remains a valid and efficient mechanism for initiating supervised classification on high-volume industrial datasets where manual labeling is not feasible (Prabhu et al., 2022).

After labeling, the text data are converted into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF). In this weighting scheme, the *Term Frequency* (TF) component captures the importance of a term within a single document. In contrast, the *Inverse Document Frequency* (IDF) penalizes terms that appear frequently across the collection. As a result, terms that generate high TF-IDF scores appear frequently in specific reviews but remain rare globally (Sheridan et al., 2025). This mechanism measures the importance of specific keywords in distinguishing individual documents from the aggregate collection, effectively strengthening the signal of discriminatory technical descriptors (such as timeout, bug, or crash) necessary for accurate error diagnosis (Sheridan et al., 2025; Shi et al., 2021). The TF-IDF matrix provides reliable input for subsequent Support Vector Machine (SVM) classification by transforming unstructured text into a high-dimensional vector space where specific error terminology carries greater weight, ensuring the model focuses on the semantic quality of complaints rather than mere word frequency.

2.4 Risk classification setup

The final phase of the diagnostic workflow implements risk measurement using Support Vector Machine (SVM) classification. The SVM was chosen as the core inference engine due to its proven reliability in handling the high-dimensional and sparse feature space characteristic of TF-IDF vectors (Nayoga et al., 2023). An optimal hyperplane that maximizes the margin between two classes is constructed by SVMs (Pinem et al., 2018). This margin maximization principle ensures that the model focuses on the most difficult data points, known as *support vectors* (Pinem et al., 2018), thereby improving its ability to generalize previously unseen user complaints. The model effectively minimizes the generalization error bound by prioritizing the separation margin, making it highly reliable in detecting critical service failures on noisy industrial datasets (Valkenborg et al., 2023).

This study specifically applies linear kernels to optimize computational efficiency without sacrificing classification performance. In the context of text mining, feature vectors generated by

TF-IDF can generally be linearly separated due to the high dimension of the vocabulary space, and linear SVMs are known to perform well when the number of features exceeds the number of samples (Mujahid et al., 2024; Thakur et al., 2025). Consequently, mapping data to a higher-dimensional space using nonlinear kernels (such as radial basis function or polynomial) can incur additional computational costs and hyperparameter tuning complexity without guaranteeing performance improvement, especially when the data are already close to linear separability (Du et al., 2024; Shamsi & Beheshti, 2025). Furthermore, in high-dimensional sparse text representations, more complex kernels can increase the risk of overfitting if the model complexity is not carefully controlled (Du et al., 2024). The system maintains a critical balance between model complexity and diagnostic speed by adhering to linear decision boundaries, which is essential for processing large-scale datasets that require rapid classification (Jayanti et al., 2025; Mujahid et al., 2024; Thakur et al., 2025).

The dataset was tested using a stratified sampling protocol with three train-test ratios: 70:30, 75:25, and 80:20 to rigorously validate the diagnostic model's stability. This multi-scenario validation approach serves as a stress test to assess the model's sensitivity to data volume variations. This comparative evaluation ensures that the reported accuracy and risk detection capabilities are consistent and robust across different configurations rather than relying on a single static split. This proves that the performance of the model is driven by learned semantic patterns related to system failures rather than simply artifacts of a particular data partition.

Furthermore, the primary objective of this study is the forensic extraction of technical failure dimensions across two national ecosystems. Therefore, this classification framework solely relies on a linear SVM model. In this specific diagnostic context, linear SVM has been empirically proven to provide the necessary balance between computational efficiency and strong separation for high-dimensional TF-IDF features (Nayoga et al., 2023), thereby eliminating the need for multi-model comparison.

2.5 Evaluation Metrics

This study uses the Confusion Matrix as a basic evaluation tool to assess the diagnostic reliability of the classification model. This matrix provides a detailed analysis of the model's predictive performance by mapping prediction labels to the actual pseudo-ground truth, identifying four different results: *true positives* (service failures identified correctly), *true negatives* (functional instances identified correctly), *false positives* (false alarms), and *false negatives* (missed failures) (Zeng, 2025). The cost of classification errors in predictive maintenance is generally asymmetric: *False negatives* (failing to detect impending failures) increase the risk of unplanned downtime and potentially fatal failures, whereas false positives primarily cause unnecessary maintenance interventions and short-term production disruptions (Raparthi et al., 2023). As a result, the confusion matrix provides a baseline from which various performance metrics are derived, enabling classification evaluation that goes beyond simple error rates and explains what the model does correctly, where it goes wrong, and how its performance can be improved for specific applications (Sathyanarayanan, 2024).

Based on the confusion matrix, the effectiveness of the model is measured by *accuracy*, *precision*, *recall*, and *F1-score*. Although *accuracy* measures the overall proportion of correctly classified instances, it can be misleading in imbalanced datasets, where classifiers that always predict the majority class can still achieve very high accuracy (Cabot & Ross, 2023). This situation often occurs in user feedback data because positive reviews far outnumber complaints. Therefore, the f1-score, the harmonic mean of precision, and recall are prioritized in this study (Zeng, 2025). A high *f1-score* indicates that the diagnostic system maintains a balanced performance: it is sufficiently sensitive to detect most technostress triggers (high *recall*, minimizing missed incidents) while maintaining a low false alarm rate (high *precision*). This balance is crucial in automated warning systems, where excessive false alarms undermine trust and missed incidents compromise the system's usability, directly affecting the effective signal-to-noise ratio of detected incidents (Vychuzhanin & Vychuzhanin, 2025).

Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values are generated to comprehensively evaluate classification performance without being influenced by specific decision thresholds. ROC curves illustrate the trade-off between *True Positive Rate* and *False Positive Rate* at various threshold settings. The associated AUC value serves as a scalar measure of the model's separation ability (Gneiting & Walz, 2021); a value close to 1.0 indicates that the model has high confidence in distinguishing between "Technostress" and "Satisfaction" events. Model validation using AUC is essential to demonstrate that the system's predictive power is stable and general rather than the result of a specific operating point or dataset split.

3. Results

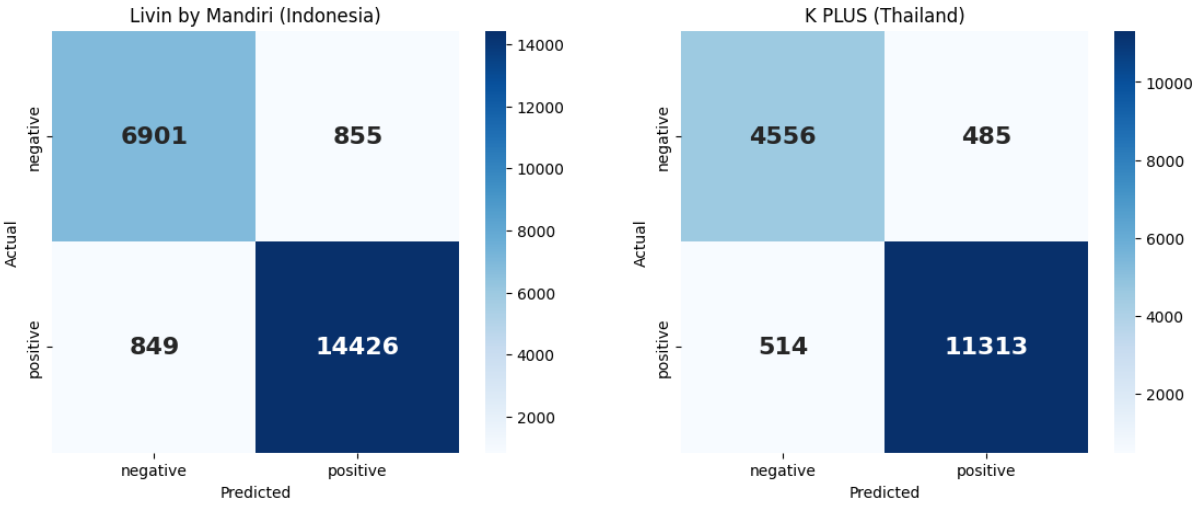
3.1 Diagnostic model reliability

The reliability of the proposed SVM classification has been rigorously verified against pseudo-ground truth to ensure the accuracy of automated forensic analysis. Table 1 summarizes the performance of the SVM model in three scenarios (70:30, 75:25, and 80:20). The reported metrics (*precision*, *recall*, and *F1 score*) are weighted averages that account for the class distribution of the dataset.

Table 2 Performance metrics of the SVM model across data split ratios

Dataset	Split Ratio	Accuracy	Precision	Recall	F1-Score
Livin' by Mandiri (Indonesia)	70:30	92.60%	92.60%	92.60%	92.60%
	75:25	92.60%	92.59%	92.60%	92.59%
	80:20	92.57%	92.55%	92.57%	92.56%
K PLUS (Thailand)	70:30	94.05%	94.07%	94.05%	94.06%
	75:25	94.08%	94.09%	94.08%	94.08%
	80:20	94.06%	94.07%	94.06%	94.07%

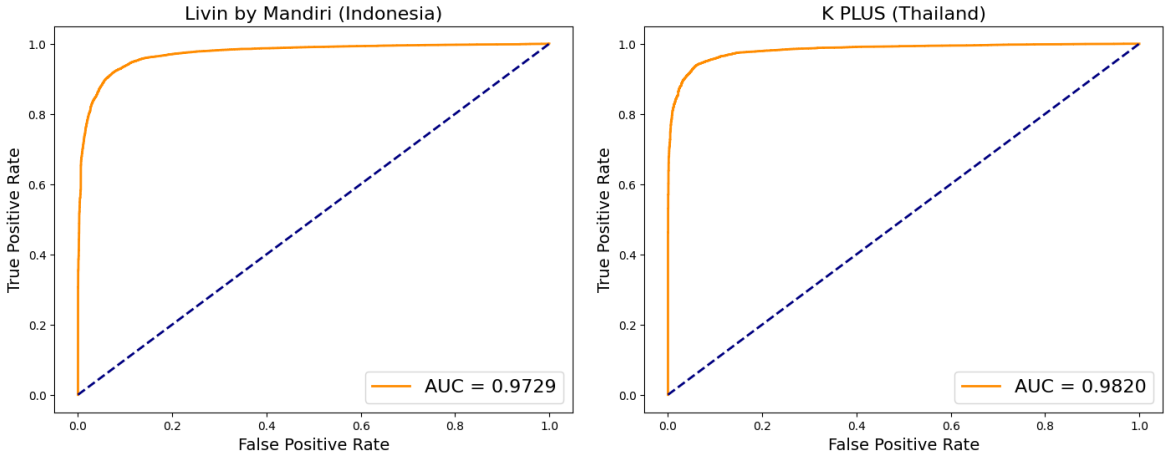
For the Indonesian dataset (Livin' by Mandiri), the model achieved an F1-score of 92.59% at the 70:30 split, while the Thai dataset (K PLUS) achieved an F1-score of 94.08% at the 75:25 split. To verify that the model is not biased toward the majority class, classification errors were visualized using a confusion matrix, as shown in Figure 2.



(a) (b)

Figure 2 Confusion matrices for Livin' by Mandiri (a) and K PLUS (b) datasets

Failure-related reviews (negative sentiment) were treated as the positive class in this study because the diagnostic objective is to detect reliability failures. The model correctly identified 6,901 failure-related reviews (True Positives) for the Livin' by Mandiri dataset, while only 855 failure reports were misclassified as satisfaction-related instances (False Negatives). Similarly, the classifier correctly detected 4,556 failure reports (True Positives) for the K PLUS dataset but missed only 485 failure-related instances. To further verify the model's robustness across decision thresholds, the Receiver Operating Characteristic (ROC) curves for both ecosystems were analyzed, as shown in Figure 3. The ROC curves for both datasets rise sharply toward the upper-left corner, yielding AUC values of 0.9729 for Livin' by Mandiri and 0.9820 for K PLUS.



(a) (b)

Figure 3 ROC curves and AUC values for the SVM-based failure classification on the Livin' by Mandiri (a) and K PLUS (b) datasets

3.2 Uncovering the Technical Dimensions

Following the probabilistic aspect extraction process using Latent Dirichlet allocation (LDA), unstructured user feedback is processed into separate latent topics. Based on the coherence score (C_v) evaluation across iterations, the optimal number of topics was determined to be $K = 3$ for the Indonesian dataset (Livin' by Mandiri) and $K = 4$ for the Thai dataset (K PLUS). Table 2 displays the aspect labels and their corresponding dominant keywords extracted using the LDA model.

On the basis of the semantic representation of the main keywords, qualitative labels were assigned to each topic to categorize service dimensions. For the Livin' by Mandiri dataset, Topic 1 was labeled Access based on prominent keywords such as "update," "time," and "open." Topic 2 is labeled 'Usability', driven by transactional and experiential terms including "transaction", "easy", and "helpful". Topic 3 is labeled Account and is characterized by terms related to banking properties such as "balance", "card", and "number". The four extracted topics for the K PLUS dataset were categorized similarly. Topic 1 was labeled "Access" and was represented by access-related keywords such as "app," "slow," and "enter." Topic 2 is labeled 'Usability' and contains positive experience terms such as "convenient", "easy", and "fast". Topic 3 is labeled Features and covers functional keywords such as "money," "account," and "card." Finally, Topic 4 is specifically labeled System Stability and is marked by critical operational keywords, including "problem," "system", "delete", and "reload."

3.3 Aspect-Based Sentiment Distribution

Mapping the sentiment polarity labels predicted by the SVM classifier onto the latent topics extracted by the LDA model measures the sentiment distribution for each technical dimension. Figure 4 illustrates the sentiment distribution by aspect, comparing the volume of positive and

negative feedback across the identified topics for both banking ecosystems.

Table 3 Latent technical dimension taxonomy in mobile banking ecosystems

Dataset	Topic	Dominant Keywords	Aspect Label
Livin' by Mandiri (Indonesia)	1	update, transfer, even, please, time, want, though, open, always, still	Access
	2	transaction, make, helpful, easy, thank, great, excellent, easier, loan, feature	Usability
	3	good, bank, balance, card, account, money, service, using, thousand, number	Account
K PLUS (Thailand)	1	app, please, update, slow, access, enter, fix, improve, wrong, press	Access
	2	good, use, convenient, easy, service, fast, much, thank, excellent, great	Usability
	3	money, bank, account, new, transfer, want, phone, number, card, apply	Feature
	4	time, bad, like, problem, system, often, day, delete, every, reload	System Stability

For the Livin' by Mandiri dataset (Figure 4), the sentiment distribution shows that positive sentiment significantly dominates the Usability and Account dimensions. In contrast, the Access dimension exhibited the highest concentration of negative sentiment, with the number of negative reviews exceeding that of positive reviews.

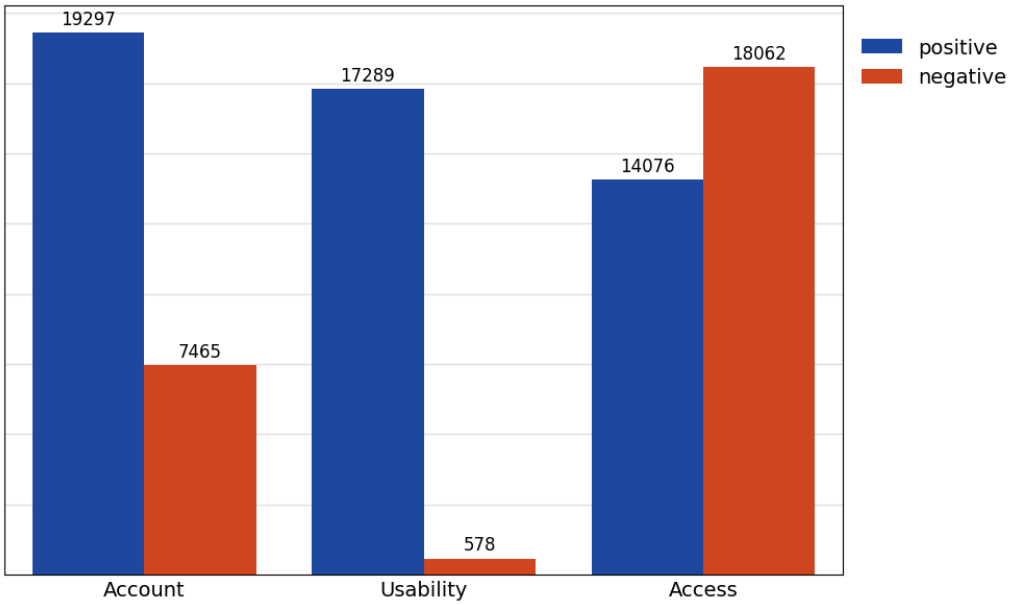


Figure 4 Aspect-Based Sentiment Distribution for the Livin' datasets

For the K PLUS dataset (Figure 5), the Usability and Features dimensions also mostly recorded positive feedback. However, in two distinct topics, negative sentiment outweighed positive sentiment: system stability and access. Specifically, the System Stability dimension recorded the highest absolute volume of negative reviews among all topics extracted from the Thailand dataset.

4. Discussion

4.1 Interpretation of failure modes

As described in Section 3.3, the sentiment distribution across topics extracted via LDA indicates differences between the two ecosystems in terms of technical challenges. Although this study does not analyze internal system logs or backend telemetry, the textual evidence collected serves as a strong indicator of users' perceived software reliability. By analyzing the co-occurrence of specific keywords and the associated sentiment polarity, the operational issues experienced by users in real-world scenarios can be inferred.

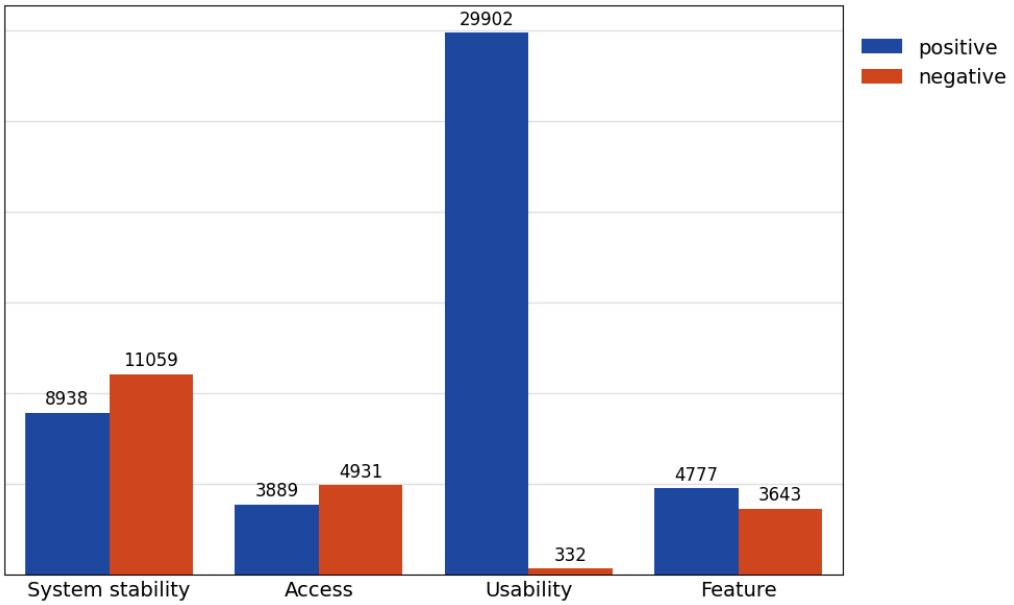


Figure 5 Aspect-Based Sentiment Distribution for the K PLUS datasets

For the Indonesian ecosystem (Livin' by Mandiri), the access dimension recorded the highest negative sentiment volume. The probabilistic clustering of keywords such as "update," "open," and "time" indicates that user dissatisfaction is concentrated at the app's entry points, particularly following software updates. The text data reveal a systemic pattern of access disruptions rather than isolated disruptions. For example, representative user feedback frequently highlighting these friction points can be found in Supplementary Material S1.

This linguistic pattern is closely linked to system degradation or congestion at access points from a data-driven perspective. Although a definitive diagnosis requires internal validation, the numerous user reports of long wait times following the update strongly suggest that users in Indonesia are largely experiencing access issues related to server-side problems or connectivity.

However, the ecosystem of K PLUS exhibits a distinct failure profile at the usage level. The emergence of system stability as the most frequently discussed topic in a negative context—marked by keywords such as "problem," "system," "delete," and "reload"—reveals a critical behavioral adjustment mechanism. The data show that users are often forced to uninstall and reinstall the app to resolve persistent runtime errors. This sentiment is reflected in the representative reviews presented in Supplementary Material S2.

Although confirming specific issues, such as memory leaks or OOM exceptions, requires a direct analysis of application failures, the systematic collection of these complaints provides a reliable external indicator. Textual evidence indicates that the client-side instability disproportionately affects users in Thailand. This symptom is consistently reported across various reviews, indicating that maintaining the application's resilience and compatibility among the highly fragmented Android device landscape remains a major technical challenge for the platform.

4.2 Security Paradox in Fintech Adoption

One of the most notable findings of the unsupervised topic extraction is the absence of data security or privacy issues as a dominant negative dimension. Although the global discourse on cybersecurity is growing, textual feedback from the Indonesian and Thai datasets largely focuses on system availability and transaction success. This empirical observation aligns with the latest e-Service Quality (e-ServQual) research by Adiningtyas and Auliani (2024), who reported that the Privacy dimension is the least discussed attribute among Indonesian mobile banking users, whereas System Availability dominates user concerns. This study highlights the existence of a systemic risk perception gap—or the security paradox—by confirming this phenomenon across two distinct national ecosystems through text mining. Textual data indicate that users of mobile banking exhibit implicit trust in the banking infrastructure. Consequently, in this context, technology-related stress is rarely triggered by fears of financial loss or data breaches; rather, such stress is almost exclusively triggered by the inability to conduct transactions due to technical failures, as identified in Section 4.1. In fact, in the analyzed dataset, security protocols generally only appear in user complaints when they act as operational barriers. Strict authentication mechanisms often cause difficulties during the login process, thereby contributing to access failures. This behavior is reflected in the representative feedback detailed in Supplementary Material S3.

Although this study does not evaluate the actual cybersecurity conditions of these applications, this user behavior has significant implications for the field of applied computer science. Users overwhelmingly prioritize convenience and service availability over visible security protocols, indicating that traditional, high-friction security measures may intensify technostress. To align with user expectations derived from this text analysis, the findings emphasize that digital banking developers should prioritize invisible or frictionless security mechanisms—such as behavioral biometrics or backend anomaly detection. This approach balances institutional needs for strict security with users' primary demands for seamless system availability.

4.3 Practical implications for reliability engineering

This study's findings offer actionable implications for mobile banking service providers. First, the proposed text mining framework demonstrates that review-based diagnostics can be used as a scalable early-warning monitoring layer. Operators can identify perceived reliability barriers and emerging sources of technology-related stress that might be missed by conventional, static surveys by systematically analyzing user feedback. Second, cross-country comparisons highlight that reliability issues are not uniform across the ecosystem; therefore, mitigation strategies must be localized. For the Indonesian market (Livin' by Mandiri), where the primary symptoms revolve around post-update login failures, the operator should prioritize smoothing the deployment process. Strengthening quality assurance protocols before application updates and ensuring entry-point gateway stability during peak traffic can significantly reduce access-related technostress reported by users. For the Thai market (K PLUS), where users frequently report app crashes and forced reloads, the focus should shift to client-side resilience. Optimizing application stability and expanding device compatibility testing to handle the highly fragmented Android ecosystem are critical steps to resolve reported runtime instability. Finally, the identified security paradox reveals that while users value secure platforms, they tolerate reduced security visibility as long as seamless service availability is maintained. To prevent security measures from becoming a source of technostress, frictionless authentication mechanisms should be prioritized in banking platforms. Implementing invisible security measures driven by backend systems—such as risk-based anomaly detection—can help minimize user friction during the login process while maintaining the necessary protection standards.

4.4 Threats to Validity

However, this study acknowledges several threats to validity.

Construct validity: Although the use of VADER-based pseudo-labels can introduce labeling bias to a certain extent, its impact is minimized by integrating Latent Dirichlet allocation (LDA) for error isolation. The diagnostic patterns inferred in this study stem not only from sentiment scores but also from the triangulation of probabilistic topic clusters and corresponding sentiment distributions. Recent evidence suggests that even with weak labels, large-scale datasets enable machine learning models, such as SVMs, to effectively capture dominant semantic patterns, provided that samples with extreme polarities are prioritized to reduce noise.

Methodological validity: This automated risk classification relies on a Support Vector Machine (SVM) architecture. Although SVM demonstrated highly reliable performance and stability for sparse, high-dimensional text representations in this study, its performance was not compared with that of alternative machine learning models. The limitation of using a single model opens opportunities for future research to explore the comparative effectiveness of models because the primary focus is on cross-country technical diagnostics rather than algorithmic optimization.

Internal validity: The automated translation of Indonesian and Thai reviews into English may introduce semantic distortion, potentially affecting polarity scoring and feature extraction. However, to preserve methodological symmetry, translation was applied consistently across both datasets. Additionally, LDA topic modeling is sensitive to preprocessing choices (e.g., stopword removal and n-gram formation), which may influence the resulting topic structure.

External validity: The analysis focuses on two market-leading applications and a single platform (Google Play Store). Therefore, the extracted failure modes may not be applicable to all banking applications or to iOS user populations. Furthermore, user reviews represent voluntary feedback and may overrepresent extreme experiences.

Despite these limitations, the study provides a large-scale, comparative diagnostic view of user complaints related to reliability through a reproducible, scalable analytical pipeline.

5. Conclusions

This study demonstrates that an automated forensic framework integrating Support Vector Machine (SVM) and Latent Dirichlet Allocation (LDA) provides a scalable diagnostic alternative for monitoring fintech reliability, replacing traditional surveys. A comparative analysis reveals two distinct modes of technical failure on user feedback: Indonesian users predominantly report post-update access barriers and authentication difficulties. In contrast, Thai users frequently complain about client-side runtime instability and application crashes. Additionally, a security paradox was identified: textual feedback indicates that users in both ecosystems prioritize seamless service availability over visible data privacy measures, highlighting the need for backend systems-driven seamless security solutions. Future research is recommended to incorporate star ratings as an additional weak label to refine sentiment monitoring to improve diagnostic granularity.

This study has several limitations. First, the diagnostic pipeline relies on weak supervision and does not use fully human-annotated ground truth labels at scale. Second, automated translation may reduce linguistic nuance in the original Indonesian and Thai texts. Third, the analysis is based solely on user-generated reviews and does not incorporate internal telemetry, such as crash logs, server monitoring metrics, or release engineering records. Future work may integrate multi-source evidence (e.g., star ratings, release timelines, and time-series volatility) to improve diagnostic granularity and strengthen causal inference regarding deployment events.

Furthermore, time-series analysis is recommended to directly correlate sentiment volatility with specific application release events (before/after updates), enabling engineering teams to measure the impact of software deployments on user stability metrics.

Acknowledgments

The authors would like to express their gratitude to Telkom University for providing the financial support and funding for this research.

Author Contributions

Arya Anggadipa Manova: Writing – Original Draft, Writing – Review & Editing, Methodology, Software, Data Curation, Visualization. Muhardi Saputra: Writing – Original Draft, Conceptualization, Validation, Supervision. Riska Yanu Fa'rifah: Writing – Original Draft, Methodology, Validation, Supervision. Sitthidej Bamrungsap: Writing – Original Draft, Investigation, Validation, Supervision.

Conflict of Interest

The authors declare no conflicts of interest.

Supplementary Materials

The following are the detailed user review excerpts supporting the extracted technical dimensions: S1. "After the update, the login process takes a long time and sometimes has to be repeated several times, even though the internet network is good and smooth" S2. "Now I'm having a problem with not being able to use the app. What is not working? Turn off the machine and turn it on again, same thing. Delete and reload and it's still the same. It doesn't work. Really bad." S3. "I can't access the app. It asks for my phone number, which I've entered. It says Send OTP, but I haven't received it since this afternoon. I really can't access the app."

Data Availability Statement

The data used in this study were collected from publicly available user reviews on Google Play Store. The full raw dataset may not be publicly shared due to platform terms of service and redistribution restrictions. Aggregated results and derived analytical outputs may be provided upon reasonable request.

References

Adiningtyas, H., & Auliani, A. (2024). Sentiment analysis for mobile banking service quality measurement. In F. Mohd (Ed.), *Procedia computer science* (pp. 40–50). Elsevier B.V. Retrieved May 4, 2026, from https://www.sciencedirect.com/science/article/pii/S1877050924003363

Alkhushayni, S., & Lee, H. (2025). Multilingual sentiment analysis with data augmentation: A cross-language evaluation in french, german, and japanese. *Information*, *16*(9). https://doi.org/10.3390/info16090806

Andrian, B., Simanungkalit, T., Budi, I., & Wicaksono, A. (2022). Sentiment analysis on customer satisfaction of digital banking in indonesia. *International Journal of Advanced Computer Science and Applications*, *13*(3), 466–473. http://thesai.org/Publications/ViewPaper?Volume=13&Issue=3&Code=IJACSA&SerialNo=56

Arief, M., & Samsudin, N. (2023). Hybrid approach with vader and multinomial logistic regression for multiclass sentiment analysis in online customer review. *International Journal of Advanced Computer Science and Applications*, *14*(12), 311–320. Retrieved December 22, 2025, from https://www.scopus.com/pages/publications/85183081033?origin=scopusAI

Bank Mandiri. (2025). *Bank mandiri awali 2025 dengan pertumbuhan sehat dan berkelanjutan.* Retrieved October 10, 2025, from https://www.bankmandiri.co.id/en/press-detail?primaryKey=448659368&backUrl=/web/guest/press

Bhatt, N., & Kothari, T. (2022). Determinants of technostress: A systematic literature review. *European Journal of Business Science and Technology*, *8*(2), 159–171. Retrieved December 21, 2025, from https://www.scopus.com/pages/publications/85147346163?origin=scopusAI

Cabot, J., & Ross, E. (2023). Evaluating prediction model performance. *Surgery*, *174*(3), 723–726. Retrieved May 4, 2026, from https://pubmed.ncbi.nlm.nih.gov/37419761/

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10). Retrieved May 4, 2026, from https://dl.acm.org/doi/full/10.1145/3507900

Dabrowski, J., Letier, E., Perini, A., & Susi, A. (2022). Analysing app reviews for software engineering: A systematic literature review. *Empirical Software Engineering*, *27*(43). Retrieved May 4, 2026, from https://link.springer.com/article/10.1007/s10664-021-10065-7

Datla, V. (2023). Article id: Ijcet_14_03_019 ensuring cloud service uptime and reliability. *International Journal of Computer Engineering and Technology (IJCET)*, *14*(3), 181–186. Retrieved May 4, 2026, from https://iaeme-library.com/index.php/IJCET/article/view/IJCET_14_03_019

Du, K., Jiang, B., Lu, J., Hua, J., & Swamy, M. (2024). Exploring kernel machines and support vector machines: Principles, techniques, and future directions. *Mathematics*, *12*(24). https://doi.org/10.3390/math12243935

Fathimath, T., & Santhi, P. (2025). Artificial intelligence driven e-services of new generation banks: Customer perception, attitude and response analysis. *International Research Journal of Multidisciplinary Scope*, *6*(3), 876–888. https://doi.org/10.47857/irjms.2025.v06i03.04268

Geeganage, D., Xu, Y., & Li, Y. (2024). A semantics-enhanced topic modelling technique: Semantic-lda. *ACM Transactions on Knowledge Discovery from Data*, *18*(4). Retrieved May 4, 2026, from https://dl.acm.org/doi/10.1145/3639409

Gneiting, T., & Walz, E. (2021). Receiver operating characteristic (roc) movies, universal roc (uroc) curves, and coefficient of predictive ability (cpa). *Machine Learning*, *111*, 2769–2797. Retrieved May 4, 2026, from https://link.springer.com/article/10.1007/s10994-021-06114-3

Jadil, Y., Rana, N., & Dwivedi, Y. (2021). A meta-analysis of the utaut model in the mobile banking literature: The moderating role of sample size and culture. *Journal of Business Research*, *132*, 354–372. https://doi.org/10.1016/j.jbusres.2021.04.052

Jayanti, L., Anam, S., Ardiyansa, S., & Maharani, N. (2025). Health insurance claim classification using support vector machine with velocity pausing particle swarm optimization. *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, *10*(2), 698–710. Retrieved May 4, 2026, from https://ejournal.uin-malang.ac.id/index.php/Math/article/view/31914

Kasikorn Bank. (2025). *Kbank unveils 2025 plans - focuses on leveraging technology, driving productivity growth and enhancing customers' experience*. Retrieved October 10, 2025, from https://www.kasikornbank.com/en/news/pages/2025_plans.aspx

Lutfi, Y., Saputra, M., & Fa'rifah, R. (2023). Aspect-based sentiment analysis in identifying factors causing technostress in fintech users using naïve bayes algorithm. *Proceedings of the International Conference on Enterprise and Industrial Systems (ICOEINS 2023)*, 107–117. https://doi.org/10.2991/978-94-6463-340-5_10

Miah, M., Kabir, M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, *14*. Retrieved May 4, 2026, from https://rdcu.be/fgMvR

Mujahid, M., Kına, E., Rustam, F., Villar, M., Alvarado, E., De La Torre Diez, I., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(87). https://doi.org/10.1186/s40537-024-00943-4

Murinde, V., Rizopoulos, E., & Zachariadis, M. (2022). The impact of the fintech revolution on the future of banking: Opportunities and risks. *International Review of Financial Analysis*, *81*. https://doi.org/10.1016/j.irfa.2022.102103

Nayoga, S., Saputra, M., & Fa'Rifah, R. (2023). An analysis of fintech technostress and its impact on smart economy: A customer review-based sentiment analysis approach. *10th International Conference on ICT for Smart Society, ICISS 2023 - Proceeding*. https://doi.org/10.1109/ICISS59129.2023.10291581

Pinem, F., Andreswari, R., & Hasibuan, M. (2018). Sentiment analysis to measure celebrity endorsment's effect using support vector machine algorithm. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 690–695. Retrieved November 4, 2025, from https://ieeexplore.ieee.org/document/8752687

Prabhu, S., Brahma, A., & Misra, H. (2022). Customer support chat intent classification using weak supervision and data augmentation. *ACM International Conference Proceeding Series*, 144–152. Retrieved May 4, 2026, from /doi/pdf/10.1145/3493700.3493729?download=true

Rahman, N., Idrus, S., & Adam, N. (2022). Classification of customer feedbacks using sentiment analysis towards mobile banking applications. *IAES International Journal of Artificial Intelligence*, *11*(4), 1579–1587. https://doi.org/10.11591/ijai.v11.i4.pp1579-1587

Raparthi, M., Dodda, S., & Maruthi, S. (2023). Predictive maintenance in manufacturing: Deep learning for fault detection in mechanical systems. *Dandao Xuebao/Journal of Ballistics*, *35*(2), 59. Retrieved May 4, 2026, from https://doi.org/10.52783/dxjb.v35.116

Sathyanarayanan, S. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, 4023–4031. https://doi.org/10.53555/ajbr.v27i4s.4345

Shamsi, M., & Beheshti, S. (2025). Separability and scatteredness (s&s) ratio-based efficient svm regularization parameter, kernel, and kernel parameter selection. *Pattern Analysis and Applications*, *28*(1), 33. Retrieved May 4, 2026, from https://link.springer.com/article/10.1007/s10044-025-01411-2

Sheridan, P., Ahmed, Z., & Farooque, A. (2025). A fisher's exact test justification of the tf-idf term-weighting scheme. *The American Statistician*, *80*(1). Retrieved May 4, 2026, from https://www.tandfonline.com/doi/full/10.1080/00031305.2025.2539241#abstract

Shi, L., Li, A., & Zhang, L. (2021). Sustainable fault diagnosis of imbalanced text mining for ctcs-3 data preprocessing. *Sustainability (Switzerland)*, *13*(4), 1–14. https://doi.org/10.3390/su13042155

Silva, C., Galster, M., & Gilson, F. (2021). Topic modeling in software engineering research. *Empirical Software Engineering*, *26*(6). https://doi.org/10.1007/s10664-021-10026-0

Subramani, V. (2025). Resilience by design: Site reliability engineering in financial platforms. *International Journal of Intelligent Systems and Applications in Engineering*, *13*(2s), 87–95. Retrieved May 4, 2026, from https://ijisae.org/index.php/IJISAE/article/view/7954

Thakur, S., Kumar Tiwari, V., & Agrawal, J. (2025). Performance analysis of linear kernel support vector machine models on real-world datasets. *Int. J. Advanced Networking and Applications*, *17*(1). Retrieved May 4, 2026, from https://www.ijana.in/archives/v17-1

The Nation Thailand. (2024). *Thailand leads southeast asia in mobile banking application usage*. Retrieved October 10, 2025, from https://www.nationthailand.com/business/banking-finance/40042041

Titiakarawongse, C., Taksin, S., Ruangsawat, J., Deeduangpan, K., & Boonkrong, S. (2024). Comparative vulnerability analysis of thai and non-thai mobile banking applications. *Journal of Cybersecurity and Privacy*, *4*(3), 650–662. https://doi.org/10.3390/jcp4030031

Tseng, C., & Koy, R. (2025). How to enter the fintech industry in southeast asia: The choice between alliance and acquisition. *Asia Pacific Management Review*, *30*(2). https://doi.org/10.1016/j.apmrv.2024.100353

UOB, PwC Singapore, & Singapore FinTech Association. (2024). *Fintech in asean 2024: A decade of innovation contents*. Retrieved October 10, 2025, from https://www.pwc.com/sg/en/publications/assets/page/fintech-in-asean-2024.pdf

Valkenborg, D., Rousseau, A., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754–757. Retrieved November 4, 2025, from https://www.ajodo.org/action/showFullText?pii=S0889540623004298

Vayansky, I., & Kumar, S. (2020). A review of topic modeling methods. *Information Systems*, *94*. https://doi.org/10.1016/j.is.2020.101582

Visa. (2024). *The future of commerce on the cusp of change visa consumer payment attitudes study 2024*. Retrieved October 10, 2025, from https://www.visa.com.ph/content/dam/VCOM/regional/ap/singapore/global-elements/documents/visa-cpa-2024-report-ipvmc.pdf

Vychuzhanin, V., & Vychuzhanin, O. (2025). Integrated approach to diagnosing complex technical systems: Experimental validation and multidimensional efficiency assessment. *Visnyk Skhidnoukrainskoho natsionalnoho universytetu imeni Volodymyra Dalia*, (5 (291)), 5–17. https://doi.org/10.33216/1998-7927-2025-291-5-5-17

Waliszewski, K., & Warchlewska, A. (2020). Attitudes towards artificial intelligence in the area of personal financial planning: A case study of selected countries. *Entrepreneurship and Sustainability Issues*, *8*(2), 399–420. https://doi.org/10.9770/jesi.2020.8.2(24)

Watanabe, K., & Baturo, A. (2024). Seeded sequential lda: A semi-supervised algorithm for topic-specific analysis of sentences. *Social Science Computer Review*, *42*(1), 224–248. https://doi.org/10.1177/08944393231178605

Yang, P., Yao, Y., & Zhou, H. (2020). Leveraging global and local topic popularities for lda-based document clustering. *IEEE Access*, *8*, 24734–24745. https://doi.org/10.1109/ACCESS.2020.2969525

Zeng, G. (2025). Invariance properties and evaluation metrics derived from the confusion matrix in multiclass classification. *Mathematics*, *13*(16). https://doi.org/10.3390/math13162609

Zimmermann, J., Champagne, L., Dickens, J., & Hazen, B. (2024). Approaches to improve preprocessing for latent dirichlet allocation topic modeling. *Decision Support Systems*, *185*. https://doi.org/10.1016/j.dss.2024.114310