

Analytical Study of Customer Review Sentiment and Education Classification as the Basis for Customer Service Management in the Blu BCA Application

Dwi Andre Vebriansyah¹, F. Danardana Murwani¹, Heri Pratikto¹

¹Universitas Negeri Malang, Indonesia

Corresponding author email: dwi.andre.2404138@students.um.ac.id

Article History: Received on December 22, 2025, Revised on January 12, 2025, Published on January 13, 2025

Abstract: This study aims to develop a sentiment and education classification model to analyze customer reviews of the Blu BCA mobile banking application in order to identify positive and negative sentiments as a basis for improving customer service management and service quality. The study employs machine learning techniques using the Naive Bayes and Support Vector Machine (SVM) algorithms. Customer review data were collected through web scraping from the Google Play Store and processed using data cleaning, tokenization, stemming, and TF-IDF weighting. Model training and evaluation were conducted using Python with the scikit-learn library. The results indicate that both Naive Bayes and SVM models are capable of effectively classifying Indonesian-language customer reviews into positive and negative sentiment categories. Performance evaluation using confusion matrix, precision, recall, and F1-score reveals differences in accuracy and effectiveness in handling linguistic characteristics of customer feedback. This study can be applied in digital banking services, customer relationship management, service quality evaluation, and financial technology analytics. The novelty of this study lies in its focused application of machine learning-based sentiment analysis on Indonesian-language reviews of a digital banking application, integrating sentiment and education classification with customer service improvement strategies to transform large-scale customer feedback into actionable insights.

Keywords: *Sentiment analysis, naive Bayes, support vector machine, customer service, Blu BCA*

A. Introduction

The rapid advancement of digitalization in the banking sector has compelled banks to deliver services that are not only fast and efficient but also centered on user experience (Abutabenjeh, S., & Jaradat, 2018). Digital banking applications have become a primary interface between banks and customers; therefore, digital service quality plays a critical role in shaping customer satisfaction, trust, and loyalty. Blu BCA, a digital banking innovation developed by PT Bank Central Asia Tbk (BCA), represents a strategic response to the evolving expectations of digitally literate customers (Gunawan, 2022). As one of the banks with strong global brand equity, BCA faces increasing pressure to ensure that its digital service quality consistently aligns with customer expectations in an increasingly competitive digital banking landscape (Adela, C., Karnila, S., & Agarina, 2024; Suyanto, 2018).

User-generated reviews on digital platforms such as the Google Play Store provide rich, real-time insights into customer perceptions, experiences, and service-related problems encountered when using digital banking applications (Owusu-Berko, 2025; Sholikhah, A. F., 2023; Sarimole, F., 2024). These reviews capture spontaneous evaluations of service performance, system reliability, responsiveness, and overall user experience. However, the massive volume of reviews, combined with the use of informal Indonesian language, slang, abbreviations, and code-mixing, makes conventional qualitative analysis inefficient and impractical (Farida, 2016). Consequently, banks often struggle to systematically translate customer feedback into concrete service improvement priorities, despite the urgency of rapid operational responses in the digital banking industry.

Machine learning based sentiment analysis and Natural Language Processing (NLP) offer an effective solution to automatically process large-scale customer feedback by classifying sentiments and extracting dominant topics from user reviews (Bharadwaj, 2023). Nevertheless, most existing studies tend to focus solely on sentiment polarity (positive or negative) without explicitly linking sentiment outcomes to structured service quality frameworks (Dewi, M., Nabila, S., & Hilaly, 2023). This limitation reduces the practical relevance of sentiment analysis results for managerial decision-making, particularly in identifying which dimensions of service quality require immediate operational intervention (Mardani et al., 2015; Hulugh & Ikeh, 2025; Hampshire, 2025 ; (Rizun et al., 2025) .

To address this gap, this study introduces an integrative approach by combining sentiment analysis and topic modeling with the SERVQUAL framework. Customer reviews of the Blu BCA application are systematically mapped onto SERVQUAL dimensions such as reliability, responsiveness, assurance, empathy, and tangibles to identify not only sentiment trends but also specific service quality gaps reflected in user experiences. This integration strengthens the study's novelty by transforming unstructured textual data into actionable service quality indicators, enabling banks to prioritize improvements based on empirically grounded customer feedback.

Furthermore, this study emphasizes concrete operational implications by translating sentiment-based SERVQUAL findings into recommendations for digital banking service enhancement, including application performance optimization, customer support responsiveness, and feature usability improvements. Academically, the research contributes to the advancement of Indonesian-language sentiment analysis by incorporating service quality theory into NLP-based analysis. Practically, the findings provide Blu BCA with a data-driven framework to improve digital service quality, enhance customer satisfaction, and sustain competitive advantage in Indonesia's rapidly evolving digital banking ecosystem.

B. Methods

This study is defined as an experimental comparative quantitative study, aligning with established practices in machine learning-based text classification research. Unlike purely descriptive studies, this research involves systematic experimentation through the comparison of two classification algorithms Multinomial Naive Bayes and Support Vector Machine (SVM) to evaluate their performance in sentiment and education classification

tasks. The experimental design enables objective assessment of model accuracy, robustness, and interpretability using standardized evaluation metrics. This approach is consistent with machine learning research that emphasizes model training, validation, testing, and comparative analysis rather than mere description of observed phenomena (Mweshi, G., & Sakyi, 2020). The data are cross-sectional, collected within a defined period, and analyzed quantitatively to ensure reproducibility and methodological rigor.

Data Sources and Collection Techniques

The research data consist of user reviews of the Blu BCA application obtained from the Google Play Store. Data collection was conducted using web scraping techniques implemented through Python libraries, enabling the extraction of both textual data and review metadata, including star ratings, review timestamps, and application version information. The inclusion of metadata addresses the limitation of text-only analysis by allowing sentiment patterns to be contextualized based on user rating behavior, temporal trends, and application update cycles, which have been shown to enrich interpretation and improve model performance in sentiment analysis studies.

Data Preprocessing

Data preprocessing was conducted to enhance data quality and ensure compatibility with machine learning algorithms. This stage involved text cleaning to remove emojis, URLs, special characters, and non-informative symbols. Indonesian stopwords were removed, slang and non-standard words were normalized into their standard forms, followed by tokenization and stemming using the Nazief-Adriani algorithm via the Sastrawi library. Subsequently, TF-IDF (Term Frequency-Inverse Document Frequency) weighting was applied to transform textual data into numerical feature vectors, enabling effective handling of high-dimensional text data during model training.

Data Labeling and Hybrid Data Labeling To address potential inconsistencies in labeling, this study adopts a hybrid data labeling approach that combines manual labeling with systematic validation procedures. Initial sentiment labels (positive and negative) were assigned manually based on predefined sentiment and education classification guidelines to preserve contextual accuracy, particularly given the informal and code-mixed nature of Indonesian-language reviews. The manual labeling process was independently reviewed by three validators a language expert, a neutral reviewer, and a data quality validator to ensure objectivity. Inter-reviewer disagreements were resolved through consensus discussions guided by the labeling criteria. This hybrid approach is methodologically justified as it balances semantic accuracy with consistency, which is critical in supervised machine learning tasks involving local language characteristics.

Dataset Feasibility Evaluation

The quality of the dataset was evaluated through assessments by three independent validators using a Likert scale. The score for each aspect was calculated using the mean formula:

$$\text{Mean} = \frac{v_1 + v_2 + v_3}{3}$$

The assessment results were converted into percentages and classified into dataset feasibility categories, namely Very Feasible (81–100%), Feasible (61–80%), Moderately

Feasible (41–60%), and Not Feasible (<40%), following Sugiyono (2017). This evaluation ensured that the dataset was suitable for further analysis.

Sentiment and Education Classification Model Development

Two machine learning algorithms Multinomial Naive Bayes and Support Vector Machine (SVM) were employed to enable comparative experimental analysis. Naive Bayes was selected for its computational efficiency and probabilistic modeling of word frequencies, while SVM was chosen for its strong performance in high-dimensional feature spaces. Model optimization was conducted using hyperparameter tuning on the validation dataset, including smoothing parameters for Naive Bayes and kernel type, C value, and gamma for SVM through grid search and cross-validation. Model performance was evaluated using accuracy, precision, recall, and F1-score derived from confusion matrices, allowing objective comparison of classification effectiveness and scalability.

SERVQUAL Mapping and Reliability Assurance

To reduce subjectivity in mapping sentiment and topics to SERVQUAL dimensions, this study operationalizes each SERVQUAL construct using explicit keyword clusters and service indicators derived from the literature. The mapping process was conducted independently by multiple coders, and inter-rater agreement analysis was applied to assess consistency across coders. This procedure strengthens the validity and reliability of SERVQUAL-based interpretation by ensuring that sentiment assignments to service quality dimensions are empirically grounded rather than purely conceptual.

Interpretation and Managerial Implications

The integration of sentiment analysis, metadata interpretation, and SERVQUAL dimensions enables the identification of service quality gaps that are both statistically significant and operationally relevant. Rather than offering general managerial insights, the findings are translated into specific recommendations, such as prioritizing responsiveness-related complaints during peak review periods, improving application reliability following version updates associated with negative sentiment spikes, and enhancing customer support features linked to assurance and empathy dimensions. This structured interpretation supports data-driven decision-making and provides Blu BCA with actionable guidance for continuous digital service improvement.

Interpretation of Results and Customer Service Implications

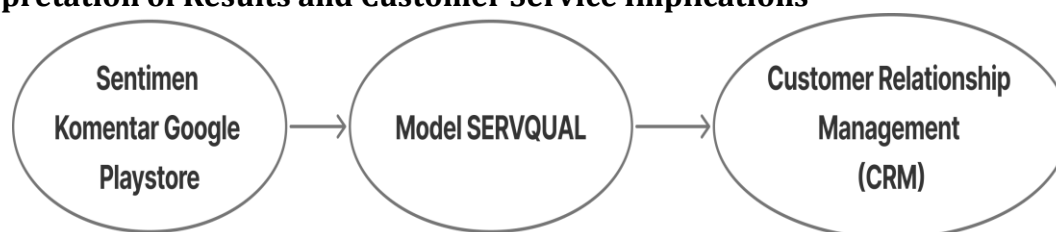


Figure 1 Conceptual Framework

The sentiment analysis results were interpreted to identify service areas requiring improvement based on the dominance of negative sentiments and the frequency of complaint-related topics. These findings were used to formulate recommendations for enhancing customer service, including prioritizing complaint handling, improving application features, and increasing service responsiveness. Furthermore, the results can

serve as a foundation for continuous customer sentiment monitoring to support data-driven decision-making at Blu BCA.

The integration of sentiment analysis results with the SERVQUAL framework provides a robust interpretative basis for evaluating the quality of customer service in the Blu BCA application. Interpretation was conducted by examining the dominance of positive and negative sentiments across each SERVQUAL dimension, along with the frequency of user complaint topics. This approach enables the identification of service areas that are not only technically problematic but also directly impact user perceptions and experiences.

C. Results and Discussion

Data Design

The data design in this study focuses on the utilization of user review texts of the Blu BCA application obtained from the Google Play Store during the period from January 2024 to August 2025. From the entire dataset collected through web crawling, only the content column was used, as it most accurately represents users' opinions and experiences. The final dataset consists of 7,235 Indonesian-language reviews, which were subsequently used as the basis for sentiment and education classification analysis employing Support Vector Machine (SVM) and Naive Bayes algorithms.

The data labeling process adopted a hybrid approach that combined automatic labeling using the Indonesian Sentiment Lexicon (InSet) with manual annotation conducted by two independent annotators. This approach was intended to improve the accuracy and consistency of sentiment labels, considering that user-generated reviews often contain informal, non-standard, and context-dependent language (Koto, F., & Rahmanningtyas, 2017; Kanapala et al., 2025; Id et al., 2024). To further ensure data reliability, an additional validator assessed the dataset quality. The evaluation results indicated a feasibility score of 78.35 percent, categorized as feasible. High scores were obtained for data relevance and sentiment label appropriateness, demonstrating that the dataset meets the required standards for subsequent sentiment analysis (Sherif, 2024; M. H. Id et al., 2024; (Zhao et al., 2023)

Data manualization or preprocessing involved several stages, including text normalization using the Colloquial Indonesian Lexicon, tokenization, stopword removal, and stemming using the Nazief-Adriani algorithm implemented through the Sastrawi library. These preprocessing steps play a critical role in reducing linguistic variation, feature redundancy, and noise within textual data, thereby enhancing the consistency and effectiveness of data representation for sentiment modeling.

Feature extraction was performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method with n-grams (1,2), followed by feature selection using the Chi-Square technique. The dataset was then divided in a stratified manner into training, validation, and testing subsets. The evaluation results demonstrate that the SVM model outperformed the Naive Bayes model in sentiment and education classification performance, indicating that SVM is more effective in capturing complex patterns in Indonesian-language user reviews. These findings confirm that systematic data design

and rigorous preprocessing significantly contribute to the success of sentiment analysis in the context of digital banking applications.

Table 1 Data Quality Validation Results

No	Aspek Penilaian	V1	V2	V3	Rata-Rata
1	Kebersihan data	3	5	3	3.67
2	Normalisasi Teks	4	4	3	3.67
3	Relevansi Data	3.5	5	4.5	4.33
4	Kesesuaian Label Sentimen	5	4	4	4.33
5	Konsistensi Labeling	4	4	4	4
6	Validasi Proses Labeling	3	5	4	4
7	Kebermaknaan Data	3	4	4	3.67
8	Kelayakan Analisis Data	4	4	3	3.67

Source: Compiled by researchers, 2025

Algorithm Implementation and Evaluation

This study implemented two machine learning-based sentiment and education classification algorithms, namely Naive Bayes and Support Vector Machine (SVM), to analyze the sentiment of user reviews of the Blu BCA application. The implementation was conducted using Python in the Google Colaboratory environment, and model performance was evaluated using accuracy, precision, recall, and F1-score metrics derived from the confusion matrix. The analysis focused exclusively on textual review data, without incorporating non-textual metadata such as numerical ratings or user profile information.

The experimental results indicate that the Naive Bayes model achieved an accuracy of 87.23 percent. The model exhibited strong performance in identifying positive sentiments, with a recall value of 94 percent, but showed limitations in detecting negative sentiments, as reflected by a recall of 68 percent. This suggests that the feature independence assumption inherent in Naive Bayes tends to make the model more sensitive to positively oriented word patterns, resulting in the misclassification of certain negative reviews (Ranataru, S., & Trianasari, 2024).

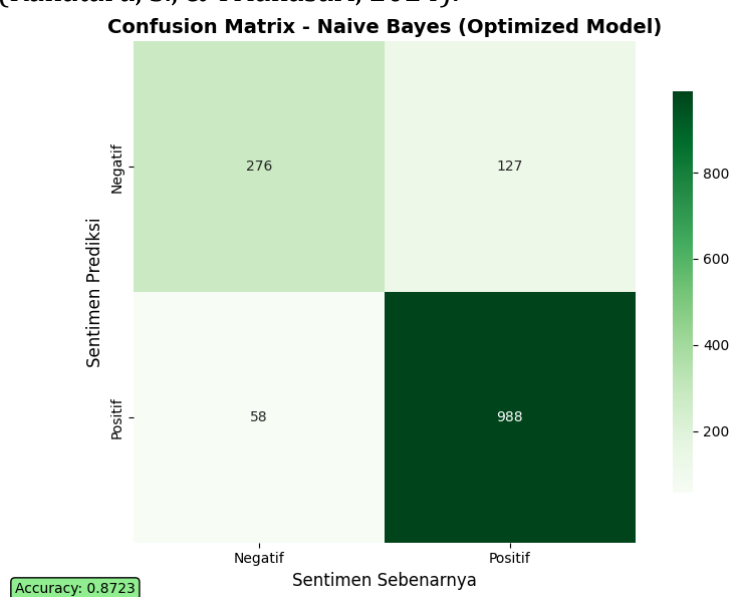


Figure 2. Naive Bayes Confusion Matrix
 Source: Compiled by researchers, 2025

In contrast, the Support Vector Machine model demonstrated superior performance, achieving an accuracy of 93.10 percent. The model produced a more balanced classification between positive and negative sentiment classes, with F1-scores of 0.95 for positive sentiment and 0.88 for negative sentiment. The lower number of misclassifications and higher confidence levels indicate that SVM is more effective in capturing complex and contextual linguistic patterns in Indonesian-language review data. Based on these results, SVM was identified as the best-performing model in this study.

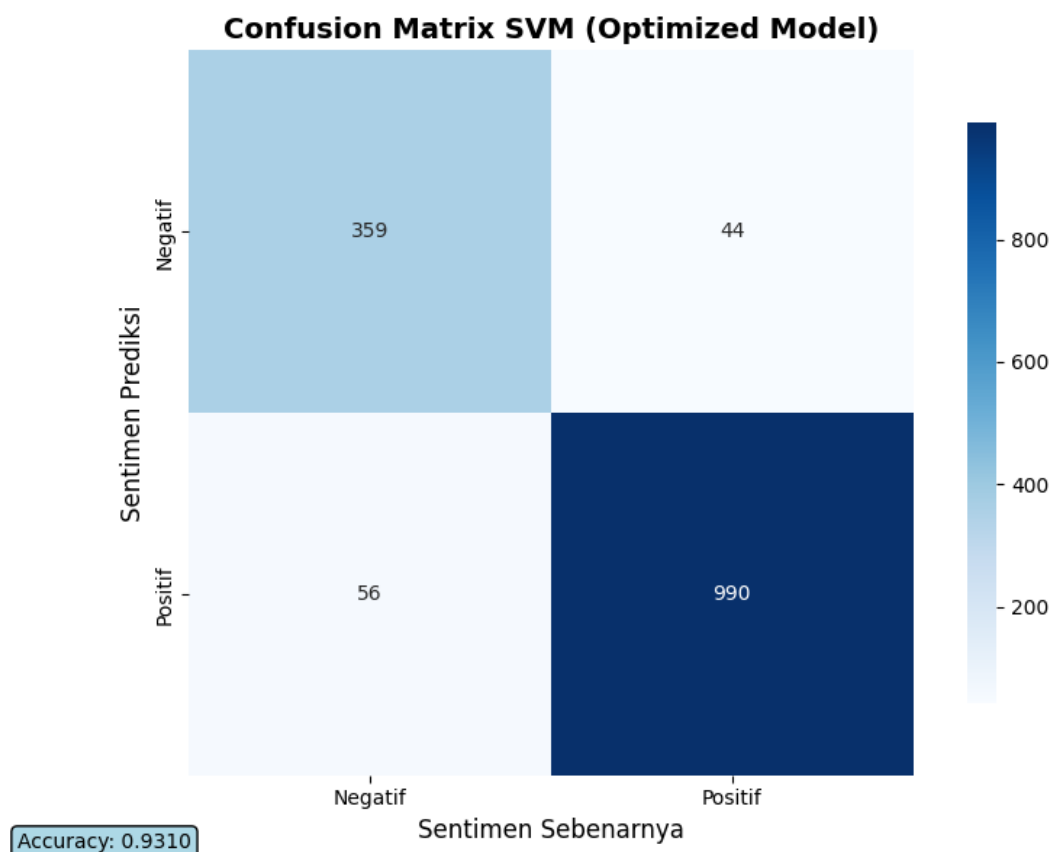


Figure 3 SVM Confusion Matrix
Source: Compiled by researchers, 2025

The sentiment and education classification results were not only used to evaluate algorithmic performance but were also integrated into the SERVQUAL framework to assess the service quality of the Blu BCA application. Each sentiment-labeled review was mapped to the five SERVQUAL dimensions Tangibles, Reliability, Responsiveness, Assurance, and Empathy based on contextual interpretation and keyword occurrence. The analysis revealed that the Reliability dimension exhibited the highest proportion of negative sentiment, primarily associated with technical issues such as application errors, login failures, bugs, and system instability (Khan et al., 2024 ; Jylhä & Hamari, 2020; Kıcıman & Fox, n.d.) The Tangibles dimension also showed a considerable level of negative sentiment related to interface design, menu layout, and overall usability (Jylhä & Hamari, 2020). Conversely, the Empathy, Assurance, and Responsiveness dimensions were dominated by positive sentiment, indicating relatively high user satisfaction with customer service attitude, perceived security, and response speed.

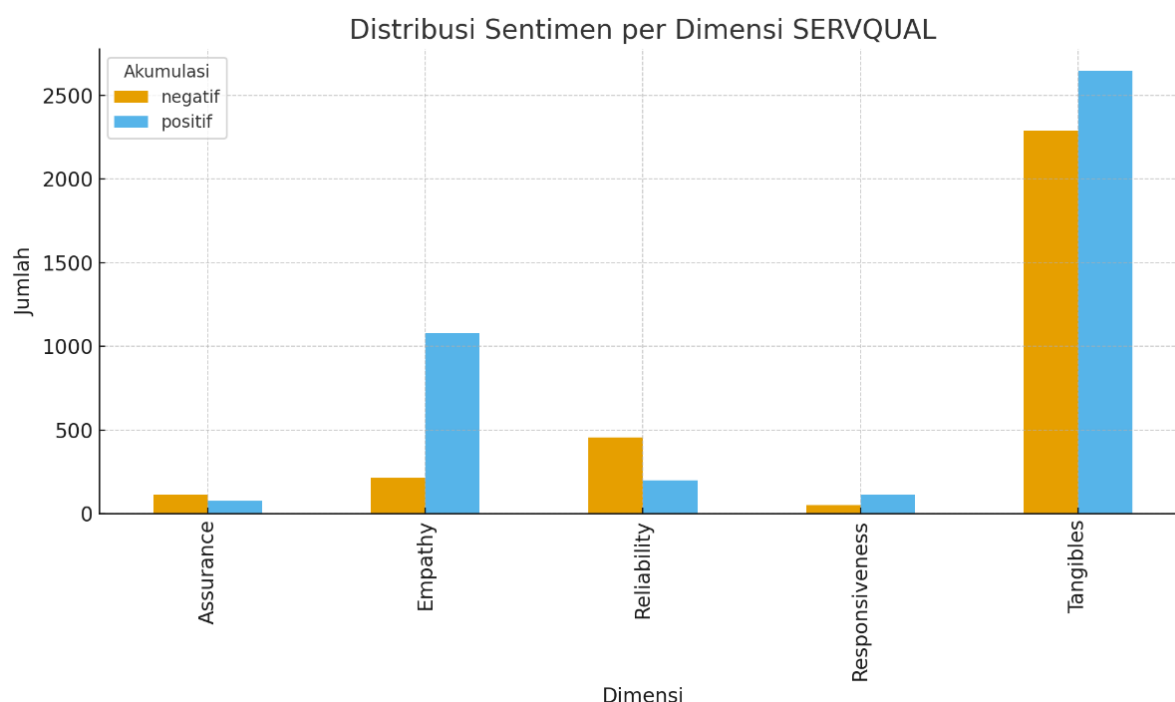


Figure 4. Sentiment Distribution per SERVQUAL Dimension

Source: Compiled by researchers, 2025

Based on these findings, service improvement efforts should prioritize enhancing system reliability and application interface quality, while maintaining and continuously improving the Empathy, Assurance, and Responsiveness dimensions to sustain user satisfaction. The integration of machine learning-based sentiment and education classification with the SERVQUAL framework provides a comprehensive analysis that not only addresses technical modeling performance but also offers valuable managerial insights to support data-driven decision-making for improving digital banking service quality.

D. Conclusion

This study concludes that the experimental comparative quantitative approach was effective in evaluating sentiment and education classification of Blu BCA application reviews using Multinomial Naive Bayes and Support Vector Machine (SVM) algorithms. The dataset demonstrated adequate quality and feasibility for analysis, supported by rigorous preprocessing and hybrid labeling procedures that enhanced data consistency and representativeness. Comparative results indicate that the SVM model outperformed the Naive Bayes model, achieving higher accuracy and more balanced performance in classifying both positive and negative sentiments, particularly in handling complex and contextual Indonesian-language expressions.

The integration of sentiment classification results with the SERVQUAL framework provided a comprehensive evaluation of digital service quality. The Reliability and Tangibles dimensions were identified as the primary sources of negative sentiment, mainly related to system stability and interface usability issues. In contrast, Empathy, Assurance, and Responsiveness were dominated by positive sentiment, reflecting higher

user satisfaction with customer service aspects. Overall, this integrated approach offers valuable, data-driven insights to support continuous improvement in digital banking service quality and managerial decision-making.

E. Acknowledgement

We would like to thank you for to the supervising lecturer who has provided direction and guidance during the research process, as well as to the university who has provided permission and facilities. The author also thanks all participants who have been willing to be respondents in this study.

References

- Abutabenjeh, S., & Jaradat, R. (2018). Clarification Of Research Design, Research Methods, And Research Methodology: A Guide For Public Administration Researchers And Practitioners. *Teaching Public Administration*, 237–256.
- Adela, C., Karnila, S., & Agarina, M. (2024). Analisis Ulasan Pengguna Aplikasi Seabank Dengan Support Vector Machine Dan Naïve Bayes. *Indonesian Journal On Software Engineering (Ijse)*, 147–155.
- Bharadwaj, L. (2023). Sentiment Analysis In Online Product Reviews: Mining Customer Opinions For Sentiment Classification. *Int J Multidiscip Res*, 5(5).
- Dewi, M., Nabila, S., & Hilaly, S. (2023). Analisis Kepuasan Pengguna Terhadap Kualitas Layanan Portal Layanan Mahasiswa (Polam) Menggunakan Metode Servqual. *Jurnal Teknologi Dan Sistem Informasi*, 1v(2), 198–202.
- Farida, T. (2016). Pengaruh Customer Relationship Management Dan Kualitas Layanan Terhadap Kepuasan Dan Loyalitas Investor Pt Danareksa Medan. *Jurnal Ilmiah Manajemen Dan Bisnis*, Xvii(1), 101–124.
- Gunawan, S. (2022). Analisis Kualitas Layanan Pengguna Website E- Learning Dengan Metode Servqual (Studi Kasus Spada Dikti Program Kampus Merdeka. *Mdp Student Conference*, Ii(2), 526–537.
- Hampshire, N. (2025). *1 St Gri Conference 2025 Impact Of Predictive Machine Learning Models On Operational Efficiency And Consumer Satisfaction In. April*, 376–403. <https://doi.org/10.63125/5tjkae44>
- Hulugh, T., & Ikeh, C. O. (2025). *Artificial Intelligence In Product Management : Automating Roadmap Prioritization Through Sentiment Analysis And Customer Feature Demand Modeling*.
- Id, M. H., Id, T. A., Id, R. I., & Id, P. U. (2024). *Leveraging Textual Information For Social Media News Categorization And Sentiment Analysis*. 1–28. <https://doi.org/10.1371/Journal.Pone.0307027>
- Id, S. K., Qasim, I., Khan, W., Khan, A., & Khan, J. A. (2024). *Plos One An Automated Approach To Identify Sarcasm In Low-Resource Language*. 1–29. <https://doi.org/10.1371/Journal.Pone.0307186>
- Jylhä, H., & Hamari, J. (2020). Development Of Measurement Instrument For Visual Qualities Of Graphical User Interface Elements (Visqual): A Test In The Context Of Mobile Game Icons. In *User Modeling And User-Adapted Interaction* (Issue 0123456789). Springer Netherlands. <https://doi.org/10.1007/S11257-020-09263-7>

- Kanapala, A., Sandeep, M., Teja, J., Manikonda, A., Mekala, B., & Mahipal, S. (2025). *Adaptive Slang And Language Modeling For Mining Informal Web Content*. 103(24), 10341–10357.
- Khan, N. D., Khan, J. A., Li, J., & Ullah, T. (2024). *Mining Software Insights : Uncovering The Frequently Occurring Issues In Low-Rating Software Applications*. <https://doi.org/10.7717/Peerj-Cs.2115>
- Kiciman, E., & Fox, A. (N.D.). *Detecting Application-Level Failures In Component-Based Internet Services*. 1–15.
- Koto, F., & Rahmaningtyas, G. (2017). *Inset Lexicon: Evaluation Of A Word List For Indonesian Sentiment Analysis In Microblogs*. 391–394.
- Mardani, A., Jusoh, A., Zavadskas, E. K., Khalifah, Z., & Nor, K. (2015). *Application Of Multiple-Criteria Decision-Making Techniques And Approaches To Evaluating Of Service Quality: A Systematic Review Of The Literature*. 16(5), 1034–1068. <https://doi.org/10.3846/16111699.2015.1095233>
- Mweshi, G., & Sakyi, K. (2020). *Application Of Sampling Methods For The Research Design*. *Archives Of Business*, 180–193.
- Owusu-Berko, L. (2025). *Harnessing Big Data, Machine Learning, And Sentiment Analysis To Optimize Customer Engagement, Loyalty, And Market Positioning*. *Int. J. Comput. Appl. Technol. Res*, 14, 1–16.
- Ranataru, S., & Trianasari, N. (2024). *Analisis Sentimen Media Sosial Terhadap Aplikasi Perbankan Untuk Mengetahui Kepuasan Pengguna Aplikasi: Studi Kasus Pada Livin By Mandiri Dan Bca Mobile*. *Al-Kharaj: Jurnal Ekonomi, Keuangan & Bisnis Syariah*, 6818–6838.
- Rizun, N., Revina, A., & Edelmann, N. (2025). *Text Analytics For Co-Creation In Public Sector Organizations : A Literature Review-Based Research Framework*. 1–45.
- Sarimole, F., & K. (2024). *Analisis Sentimen Terhadap Aplikasi Satu Sehat Pada Twitter Menggunakan Algoritma Naive Bayes Dan Support Vector Machine*. *Jurnal Sains Dan Teknologi*, 783–790.
- Sherif, S. M. (2024). *Lexicon Annotation In Sentiment Analysis For Dialectal Arabic : Consensus Expert Standardized Criteria*. 2024, 165–172.
- Sholikhah, A. F., & H. (2023). *Pengaruh Kualitas Layanan, Kualitas Produk Dan Harga Terhadap Loyalitas Pelanggan Melalui Kepuasan Pelanggan Mie Gacoan Di Bekasi Timur*. *Jurnal Economina*, 11(2), 692–709.
- Suyanto. (2018). *Machine Learning Tingkat Dasar Dan Lanjut*. [Online] *Informatika*.
- Zhao, S., Hong, X., Yang, J., Zhao, Y., & Ding, G. (2023). *Toward Label-Efficient Emotion And Sentiment Analysis*. *Proceedings Of The Ieee*, 111(10), 1159–1197. <https://doi.org/10.1109/Jproc.2023.3309299>