DECODE: JURNAL PENDIDIKAN TEKNOLOGI INFORMASI
P-ISSN: 2775-2984 | E-ISSN: 2775-1813
Vol. 6, No 1, Maret 2026, Hal: 40-50
Doi: http://dx.doi.org/10.51454/decode.v6i1.1565
https://journal.umkendari.ac.id/index.php/decode
This work is licensed under a CC BY-SA license.

# Sentiment Analysis of *Wargaku Surabaya* Apps Reviews using Naïve Bayes and Support Vector Machine (SVM) Methods

**Li'izah Nur Fadilah[1], Anik Vega Vitianingsih[1*], Yudi Kristyawan[1], Anastasia Lidya Maukar[2], Nina Kurnia Hikmawati[3]**

[1]Informatics Department, Universitas Dr. Seotomo, Indonesia.
[2]Industrial Engineering Department, President University, Indonesia.
[3]Information System Department, Universitas Komputer Indonesia, Indonesia

**Article Info**

**Keywords:**

Naïve Bayes
Natural Language Processing
Sentiment Analysis
Support Vector Machine
Text Mining
*Wargaku Surabaya* Reviews

***Article History:***

Submitted: December 15, 2026
Accepted: January 27, 2026
Published: January 27, 2026

**Abstract:** Digital public service applications in Indonesia are increasingly used to improve citizens' access to government services, generating large volumes of feedback that are difficult to analyze manually. Moreover, many previous studies focus on polarity-based sentiment, which may not adequately capture specific user emotions. This study analyzes feedback on the Wargaku Surabaya application by classifying emotions into five categories: anger, disappointment, sadness, pride, and happiness. A total of 1,406 texts were collected (2021–2025), with 1,386 retained after preprocessing. Data were primarily sourced from Google Play Store reviews, supplemented by comments from Threads and YouTube. The research employs text preprocessing, TF-IDF weighting, and lexicon-based labelling with the generated labels reviewed on a subset of the dataset before model training. Emotion classification was performed using Naïve Bayes (NB) and Support Vector Machine (SVM), evaluated via a train–test split and confusion matrix. Results show that SVM achieved 84% accuracy, 85% precision, 84% recall, and an 84% F1-score, outperforming NB with 58% accuracy. These findings indicate that SVM is more reliable for multi-class emotion classification in digital public services.

**Corresponding Author:**
Anik Vega Vitianingsih
Email: vega@unitomo.ac.id

## INTRODUCTION

As public services in Indonesia become more digital, some local governments have launched mobile apps to deliver information, reporting, and communication more efficiently. The Surabaya City administration created the *Wargaku Surabaya* app to improve communication with residents (May & Fanida, 2022). As user adoption grows, online reviews offer extensive insights into user experiences. However, the large quantity and varied nature of the feedback make manual analysis inefficient, and polarity-based sentiment analysis may fail to capture the specific emotions behind users' responses to public service quality (Nur et al., 2024). Emotion analysis improves the understanding of user feedback by identifying specific emotional responses, allowing more targeted service adjustments than polarity-based sentiment classification (Motger et al., 2025).

The *Wargaku Surabaya* application was designed to improve user experience. However, user reviews often mention recurring problems such as technical issues, disorganized data, account verification, and report submission. Most of these concerns came from Google Play Store reviews, with additional comments found on Threads and YouTube. (Atmaja et al., 2025). Other programs, like JMO, have also faced complaints about technical issues and poor user experience (Rizaldi et al., 2023). These problems highlight the need to carefully evaluate public opinion so providers can keep improving the application's quality.

Most previous studies on digital public service apps have focused on sentiment analysis rather than emotion analysis. For instance, SIGNAL research (Kacung et al., 2024), WhatsApp (Aida Sapitri & Fikry, 2023), and Mobile JKN (Maulida et al., 2024) groups' opinions are neutral, negative, or positive. This method does not identify specific emotions such as anger, disappointment, grief, pride, or happiness. Knowing these emotions could help show which parts of the service cause certain reactions. However, there is still limited research on multi-class emotion analysis for Indonesian e-government application reviews, especially studies that combine feedback from several online sources.

Recent advances in natural language processing (NLP) have enhanced the accuracy of emotion analysis and established NLP as a vital tool for analyzing unstructured textual data in research (Andana et al., 2023). This study applies standard NLP pre-processing and machine learning techniques to classify emotions in user reviews (Rifaldi et al., 2023). Naïve Bayes (NB) and Support Vector Machines (SVM) are often used for text categorization because they work well in modelling tasks and can handle high-dimensional data efficiently (Artanto, 2024). Previous research on sentiment and emotion classification has utilized a variety of methods, including classical machine learning algorithms and deep learning models. Fine-grained emotion extraction from mobile app reviews has been shown to yield more detailed insights into user experience and service quality issues (Motger et al., 2025). Furthermore, deep learning methods such as Long Short-Term Memory (LSTM) networks have been investigated in conjunction with Support Vector Machines (SVM) for analyzing user reviews, demonstrating the applicability of multiple modeling strategies in this domain (Damayanti et al., 2024). This study compares Naive Bayes (NB) and Support Vector Machine (SVM) algorithms to evaluate their effectiveness in multi-class emotion classification of digital public service feedback.

This study introduces an emotion analysis framework to examine user feedback on the *Wargaku Surabaya* app, available on the Google Play Store, YouTube, and Threads, thereby addressing a previously identified research gap. The research uses a hybrid method that involves text pre-processing, TF-IDF weighting, lexicon-based automatic labelling, and classification with Naïve Bayes and SVM algorithms. This study focuses on five emotion categories, utilizes a multi-platform dataset, and combines machine learning with lexicon-based methods to support emotion recognition in the context of Indonesian digital public services. The findings are expected to provide the Surabaya City Government with a stronger basis for evaluating its services through emotional indicators and to support the development of more user-focused, responsive digital public services.

## METHOD

To enhance the quality of analytical insights derived from user feedback, this research applies Naïve Bayes (NB) and Support Vector Machine (SVM) as classification models for emotion detection in reviews of the *Wargaku Surabaya* application. Research process stages are depicted in Figure 1.

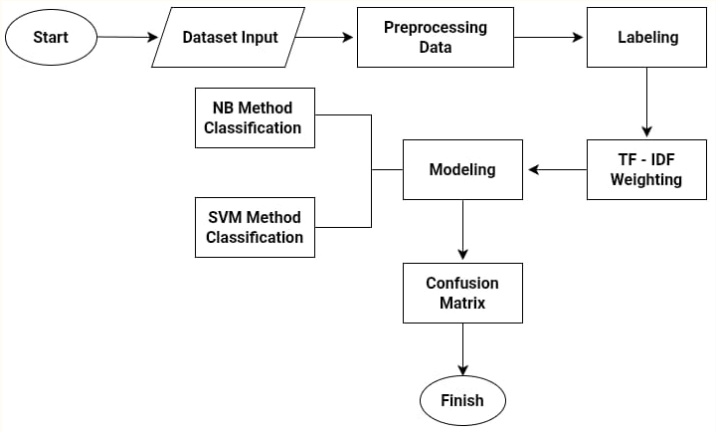


Figure 1. Research Phase

Text cleaning, case folding, tokenization, stopword removal, normalization, and stemming are part of the preparation step of the procedure, which begins with the acquisition of user review datasets from various digital platforms. The primary dataset consisted of Google Play Store reviews. Additionally, public comments on Threads and YouTube that explicitly referenced the Wargaku Surabaya application served as supplementary sources of feedback. To facilitate comparability across platforms, the collected texts were compiled into a unified dataset and processed using a standardized natural language processing (NLP) pipeline. After pre-processing, a lexicon-based method automatically labeled the reviews into five emotion categories: anger, disappointment, sadness, pride, and happiness. To ensure label reliability, the automatically generated labels were verified on a subset of the dataset before use in training supervised classification models. TF-IDF weighting is then used to convert the labelled information into numerical feature representations before processing in the modelling step. At this point, emotions are classified using TF-IDF features and the NB and SVM algorithms. The dataset was divided into training and test sets using a 75:25 ratio, with the random_state parameter set to 42. The last stage is to assess the model's performance by generating accuracy, precision, and recall values using a confusion matrix.

### Data collection

A total of 1,406 user feedback texts were collected between 2021 and 2025, primarily from Google Play Store reviews as the main data source and supplemented by publicly available comments from Threads and YouTube. Data collection was performed using Python-based web scraping with relevant APIs and scraping libraries for each platform. Following data cleaning and preprocessing, 1,386 reviews were retained and used for subsequent analysis.

### Data Preprocesing

Data pre-processing aims to simplify the dataset, maintain linguistic consistency, and remove elements that may hinder accurate emotion classification (Motger et al., 2025). At this point, raw user reviews are converted into more structured text using steps such as cleaning, case folding, tokenization, normalization, stopword removal, and stemming (Rizaldi et al., 2023).

1. *Cleaning:* This part involves refining text data by eliminating elements such as URLs, emojis, punctuation marks, numbers, and other irrelevant characters that do not contribute to the analysis (Rifaldi et al., 2023).
2. *Case Folding:* This proccess changes all characters to lowercase to keep words consistent and avoid different versions of the same word (Rifaldi et al., 2023).

3. *Tokenization:* This process breaks text into smaller components known as tokens, typically individual words, to enable more precise linguistic processing and computational analysis (Rifaldi et al., 2023).
4. *Normalization:* Normalization changes informal or misspelt words into standard Indonesian. This process helps limit the variety of words often found in user content (Andana et al., 2023).
5. *Stopword Removal:* Stopword removal takes out common words that do not add much meaning, so the model can pay more attention to words that matter for emotion (Rifaldi et al., 2023).
6. *Stemming:* This process transforms words into their base forms by removing prefixes and suffixes, which reduces term variations and yields more consistent features (Atmaja et al., 2025).

**Data Labeling**

The system automatically labels the processed review text using an emotion lexicon that groups words into specific emotional categories (Putri & Muthia, 2024). Each word in the review is checked against the lexicon for five emotional classes: anger, disappointment, sadness, pride, and happiness. When a word matches, it adds to the emotional score for that category. After matching all words, the system totals the scores for each class and assigns the review to the emotion with the highest score. This method uses the lexicon-based scoring formula shown in Equation (1).

$$L\ (e) = \frac{1}{n} \sum_{i=1}^{n} I = (w_i\ \epsilon\ E) \quad (1)$$

Here $L\ (e)$ denotes the total lexicon-based score for emotion category $w_i\ \epsilon\ E$ represents the weight assigned to the $i$-th word in the review, and $E$ refers to the predefined emotion lexicon. The value $n$ indicates the total number of words in the review that match entries in the emotion lexicon. Each review is classified into the emotion category with the highest accumulated score. The lexicon-based labelling method enables automatic annotation without large manually labelled datasets, making it useful when labelled data is scarce (Putri & Muthia, 2024). It also improves interpretability, as each emotion label is drawn directly from the lexicon, allowing researchers to see which words influence decisions (Motger et al., 2025). These methods yield deeper insights into user reactions, especially in digital services where emotions matter (Atmaja et al., 2025).

**TF-IDF Weighting**

TF-IDF is a way to measure how important a word is in a given document relative to a set of documents. It is commonly used in text classification because it emphasizes terms that are more prominent in a particular document than in others, aiding categorization and distinguishing documents (Faisal et al., 2022). When a term occurs more frequently within a document, its TF value increases. Conversely, if the term appears across numerous documents, its IDF value decreases, resulting in a lower overall weight (Purnamasari et al., 2023). TF-IDF consists of two parts: Term Frequency (TF), representing how often a term occurs, and Inverse Document Frequency (IDF), which reflects the rarity of the term across all documents. As shown in Equation (2).

$$W_{t,d}\ = tf_{t,d}.\log(\frac{N}{df_t}) \quad (2)$$

**Classification Model**

The classification model uses the Naïve Bayes and SVM methods to analyze the review data.

**1. Naïve Bayes**

Naïve Bayes (NB) is a probabilistic classifier commonly used in text analysis due to its simplicity, computational efficiency, and strong performance in high-dimensional feature spaces, such as TF-IDF (Komarudin & Hilda, 2024). The NB model estimates the likelihood that a document belongs to a class based on its words using Bayes' theorem. It speeds up calculation and is appropriate for huge text datasets because it assumes independent features (Purnamasari et al., 2023). Following feature extraction using TF-IDF, user reviews are sorted into five emotion groups using NB. Provides the primary probability function for the NB classifier, which explains how classification is determined. This process is described in Equation (3)

$$P(C_i|X) = \frac{P(X|C_i)\ X\ P(C_i)}{P(X)} \quad (3)$$

The NB algorithm assumes that features are independent. To find the probability that all features appear in a review, you multiply the probability of each feature, as shown in Equation (4).

$$P(X|C_i) = \prod_{j=1}^{n} P(x_i|C_i) \quad (4)$$

This approach simplifies the probabilistic calculation process, enabling NB to work well with high-dimensional data such as TF-IDF. As a result, it is suitable for classifying emotions in *Wargaku Surabaya* app user reviews.

**2. SVM**

A Support Vector Machine (SVM) is a supervised classification method that separates data into distinct classes by determining an optimal hyperplane in the feature space. The boundary is obtained by maximizing the class margin via an optimization procedure involving Lagrange multipliers, thereby enhancing the generalization capability of SVMs to new data (Damayanti et al., 2024). SVM is often used for text classification tasks such as emotion detection. In this study, SVM is used to sort user reviews of the *Wargaku Surabaya* app into five emotion categories: anger, disappointment, sadness, pride, and happiness. Before classification, the text is converted into TF-IDF features. The SVM decision function, shown in Equation (5), measures how an input feature vector relates to its predicted class.

$$f(x) = w.x + b \quad (5)$$

This Equation uses the feature vector to represent the TF-IDF version of the review text, the weight vector to determine the direction of the separating hyperplane, and the bias term to shift the hyperplane within the feature space. The decision function's sign determines the class to which a review belongs, helping SVM distinguish among various emotions in complex textual data.

**Evaluation Model**

The primary evaluation method employed in this research is the Confusion Matrix, which measures the accuracy of the NB and SVM models in categorizing user reviews into five emotional categories: pride, happiness, sadness, anger, and disappointment. The confusion matrix, a popular technique in multi-class classification, offers a succinct summary of accurate and inaccurate predictions for each emotion label, making it a useful tool for assessing model performance (Sathyanarayanan & Tantri, 2024). In multi-class emotion recognition, the confusion matrix is useful because it shows where the model makes mistakes and how well it distinguishes among emotions (Riehl et al., 2023). In this matrix, we can calculate performance measures such as accuracy, precision, recall, and F1-score can be calculated. These metrics show how often the model is correct, how reliable its emotion labels are, and how well it finds real emotional cases, while the F1-score provides a balanced evaluation by combining precision and recall (Zeng, 2025).

Accuracy shows the percentage of emotion predictions the model gets right across the whole dataset, as shown in Equation (6). Precision shows how many reviews predicted to be in an emotion category are actually correct, as said in Equation (7). Recall measures how well the model finds real emotional cases by comparing the number of correct predictions to the total true cases for each emotion, as shown in Equation (8).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

A true positive (TP) is when a review is correctly identified as showing a certain emotion. A false positive (FP) occurs when a review is incorrectly labelled as expressing an emotion it does not. A false negative (FN) is when a review that should be in a certain emotional category is put in the wrong one. A true negative (TN) means a review does not belong to a specific emotional category and is correctly recognized as such.

## RESULTS AND DISCUSSIONS

The dataset employed in this research was obtained by scraping user reviews of the *Wargaku Surabaya* application. A total of 1,406 reviews were collected from Google Play Store, Threads, and YouTube between 2021 and 2025. After data cleaning and preprocessing, 1,386 reviews remained for analysis.. Table 1 summarizes the web scraping results. All retrieved review data were systematically organized and stored in CSV format for subsequent pre-processing and analysis.

Table 1. Web Scraping Result

| Date | Content |
|---|---|
| *10/03/2025* | *Password salah terus, padahal sudah sesuai.* |
| *09/15/2025* | *Masih banyak sekali masalah di aplikasi nya, perlu banyak sekali perbaikan dan peningkatan* |
| *07/22/2025* | *Sangat membantu setiap permasalahan dari warga* |
| *06/23/2025* | *Aplikasi gak beres gak bisa buat laporan* |
| *04/04/2025* | *Sungguh menghambat pernikahan* |
| *04/12/2025* | *aplikasi yang sangat ngebug, kurang dari maps nya, pelayanan nya chatting nya sangan buruk* |
| *5/17/2025* | *Pak soal apk wargaku pelapor data dirahasiakan, percuma ngelapor tapi di musuhi 1 rt atau rw. Kalau bisa yang buka laporan cuma anda aja, kalau anak buah anda pas sidak tertib semua.* |

The reviews included noise and discrepancies after data collection, making them inappropriate for immediate processing. To solve these problems, a pre-processing step was implemented to convert the unprocessed text into a more organized, cleaner format suitable for emotion classification. Text cleaning, case folding, tokenization, stopword elimination, normalization, and stemming were among the pre-processing procedures. The outcomes of these pre-processing procedures are shown in Table 2

Table 2. Pre-processing Results

| | **Raw Dataset** |
|---|---|
| *Aplikasi aneh, buat daftar akun tamplate format tgl lahir (hari-bulan-tahun), tapi pas disuruh validasi mintanya (tahun-bulan-hari). kan kami ga bisa ngetik sebebas itu ya, aplikasi aneh ga jelas* | |
| | **Text pre-processing** |
| Cleaning | *Aplikasi aneh buat daftar akun tamplate format tgl lahir hari bulan tahun tapi pas disuruh validasi mintanya tahun bulan hari kan kami ga bisa ngetik sebebas itu ya aplikasi aneh ga jelas* |
| Case Folding | *aplikasi aneh buat daftar akun tamplate format tgl lahir hari bulan tahun tapi pas disuruh validasi mintanya tahun bulan hari kan kami ga bisa ngetik sebebas itu ya aplikasi aneh ga jelas* |
| Tokenisasi | *aplikasi, aneh, daftar, akun, tamplate, format, tgl, lahir, tapi, pas, disuruh, validasi, mintanya, ga , ngetik, sebebas, aplikasi, aneh, ga* |
| Stopword Removal | *aplikasi, aneh, daftar, akun, tamplate, format, tgl, lahir, tapi, pas, suruh, validasi, minta, ga, ngetik, bebas, aplikasi, aneh, ga* |
| Stemming | *aplikasi, aneh, daftar, akun, tamplate, format, tgl, lahir, tapi, pas, suruh, validasi, minta, ga, ngetik, bebas, aplikasi, aneh, ga.* |

After pre-processing, the data were labeled using a predefined emotion lexicon and a lexicon-based method. Each review was assigned to one of five emotion categories: anger, disappointment, sadness, pride, or happiness. Table 3 presents the results of the labelling process.

Table 3. Labeling Result

| Text_clean | Sentiment |
|---|---|
| *aplikasinya banyak akses ke pemerintahan tapi sayangnya muter tahu error* | *disappointment* |
| *Aplikasi kok kesalahan sistem terus ini gimana* | *sadness* |
| *Aplikasi simpel, layanan bagus, semoga kedepannya tetap konsisten, saran saya, Semoga kedepannya rt/rw& kelurahan di edukasi dalam sistem ini agar lebih responsif* | *happiness* |
| *Kacau , jujur saja ini aplikasi bukannya mempermudah justru malah mempersulit sekali* | *anger* |
| *Mantap respon cepat* | *pride* |

Once emotion labelling was complete, the review dataset was sorted into categories to identify which terms appeared most often in each emotional group. These term frequencies were displayed in word clouds to clarify the common vocabulary. Each emotion group has its own set of frequent words. For instance, in the *kecewa* (disappointed) category, words like *"daftar"* (register), *"tidak"* (not), *"gak"* (cannot), *"tapi"* (but), and *"aplikasi"* (application) are the most common, as shown in Figure 2. This suggests that users often express disappointment about registration, verification problems, and how the application works. Figure 3 presents that the words "error", "*ribet*" (complicated), "*susah*" (difficult), and "*buka*" (open) appear most often in the angry category.

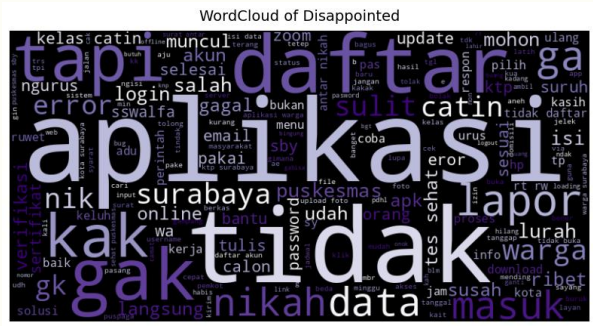


Figure 2. Disappointed Word Cloud


Figure 3. Angry Word Cloud

Figure 4 presents that the words such as "*kak*" (older sibling), *"mohon"* (please), *"tolong"* (help), and *"gimana"* (how) are most common in the sad category. Figure 5 shows that the words "Surabaya", "*mantap*" (great), "*warga*" (citizens), and "*aplikasi*" (application) are the most common in the proud category.

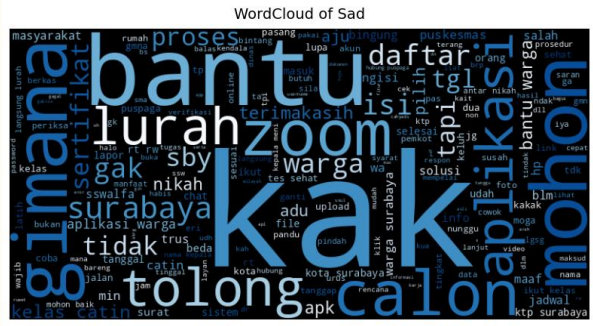


Figure 4. Sad Word Cloud

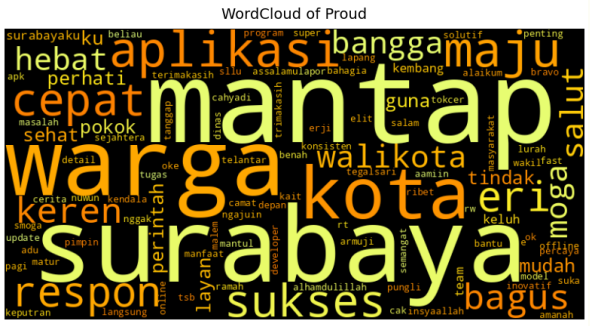


Figure 5. Proud Word Cloud

Figure 6 shows that words like *"bagus"* (good), *"bantu"* (help), *"warga"* (citizens), and *"terima kasih"* (thank you) are most common in the happy category.


Figure 6. Happy Word Cloud

Text features were extracted using term frequency-inverse document frequency (TF-IDF) weighting in Python, and a Naïve Bayes classifier was subsequently applied to categorize the reviews into five emotion categories. After preprocessing, the dataset comprised 1,386 reviews. For model evaluation, the data were partitioned into training (75%, n = 1,039) and testing (25%, n = 347) sets using a train–test split approach. Consequently, the confusion matrix results pertain exclusively to the test set. As shown in Figure 7, the Naïve Bayes model achieved the highest performance in the disappointment category, with 141 correct predictions. Nevertheless, a substantial number of reviews labeled as sad (57) and happy (25) were misclassified as disappointed, indicating a tendency of the model to over-predict the disappointment class. The model demonstrated moderate accuracy in identifying sadness and happiness, with 39 and 17 correct predictions, respectively. Predictions for anger and pride were lower, suggesting that these categories were more challenging for the model to distinguish.

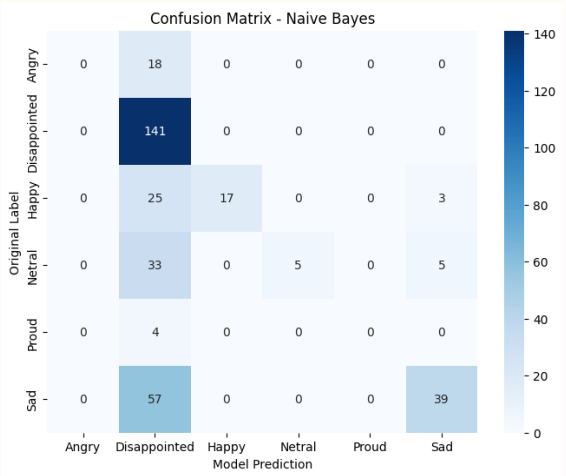


Figure 7. Naïve Bayes Classification Results

Figure 8 shows that the SVM model correctly classified several emotions, identifying 134 disappointed, 83 sad, and 33 happy reviews. The model also identified 36 neutral cases. However, the model struggled with anger and pride, frequently misclassifying them.

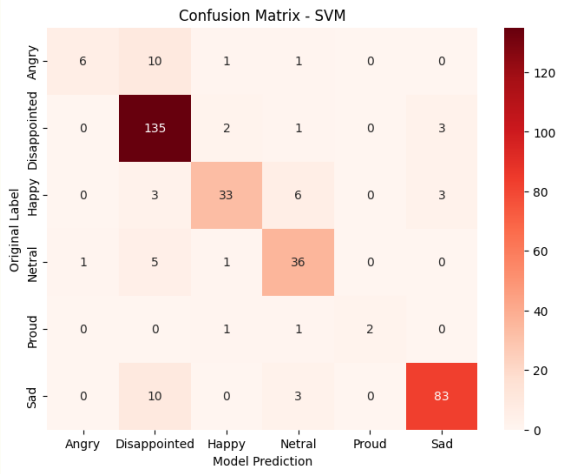


Figure 8. SVM Classification Results

Once the testing phase is complete, the next phase is the model's performance evaluation. This evaluation uses confusion matrices to compute accuracy, precision, and recall, as shown in Table 4.

Table 4. Confusion Matrix Result

| **Confusion matrix result** | | | |
|---|---|---|---|
| | Accuracy | Precision | Recall |
| NB | 58% | 69% | 58% |
| SVM | 84% | 85% | 84%, |

The confusion matrix and evaluation metrics in Table 4 show that the SVM model outperformed Naïve Bayes in classifying five emotions in *Wargaku Surabaya* App reviews. This result is consistent with earlier studies, which found that SVM performs better with high-dimensional text features such as TF-IDF and provides clearer class separation than Naïve Bayes (Aprinastya et al., 2024). In this study, the SVM algorithm reached 84% accuracy, 85% precision, 84% recall, and an F1 score of 84%. The SVM results are higher than those of Naïve Bayes, which achieved 58% accuracy, 69% precision, 58% recall, and an F1-score of 52%. These results demonstrate that SVM outperforms Naïve Bayes, achieving 84% accuracy and an F1-score of 84%, compared to 58% accuracy and a 52% F1-score for Naïve Bayes. making it a better choice for multi-class emotion classification on the *Wargaku Surabaya* dataset. The outcomes of the word cloud visualization used to analyze the emotions of *Wargaku Surabaya* users are presented in Figure 9 This graphic depicts the important terms that are commonly stated in user reviews from Threads, YouTube, and the Google Play Store, providing a better understanding of the emotional reactions users have when using the app.

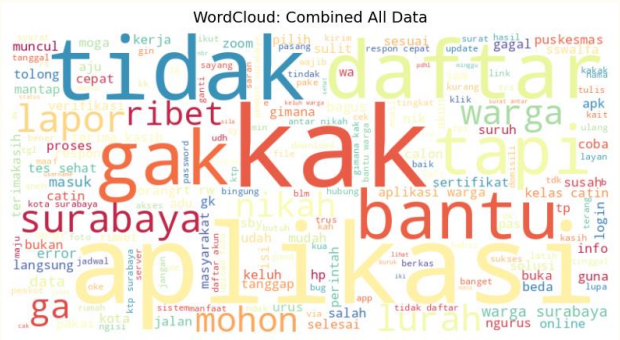


Figure 9. Word Cloud Data

## CONCLUSIONS

The application of multi-class emotion classification to user feedback, primarily sourced from Google Play Store reviews and supplemented by public comments from Threads and YouTube, yielded 1,406 texts spanning 2021 to 2025. Following data cleaning and preprocessing, 1,386 reviews were retained for analysis. Employing a lexicon-based labeling method and TF-IDF feature weighting, the reviews were categorized into five emotion categories: anger, disappointment, sadness, pride, and happiness.

The analysis revealed that negative emotional responses, particularly disappointment and sadness, were most prevalent and were frequently associated with registration difficulties, validation failures, and application errors. In the comparative model evaluation, SVM demonstrated superior performance compared to Naïve Bayes, achieving 84% accuracy, 85% precision, 84% recall, and an F1-score of 84%. By contrast, Naïve Bayes achieved 58% accuracy, 69% precision, 58% recall, and a 52% F1-score. These results indicate that SVM offers more reliable performance for multi-class emotion classification within Indonesian digital public services. Future research should expand the dataset, explore deep learning approaches, and investigate richer feature representations, such as contextual embeddings or multimodal information, to further enhance classification performance.

## REFERENCES

Aida Sapitri, I., & Fikry, M. (2023). Pengklasifikasian Sentimen Ulasan Aplikasi WhatsApp Pada Google Play Store Menggunakan Support Vector Machine. *Jurnal TEKINKOM*, *6*(1), 1–7. https://doi.org/10.37600/tekinkom.v6i1.773

Andana, M. H., Daffa, M., Fitria, N. U., & Mujiastuti, R. (2023). Webinar & Workshop Natural Language Processing in the Life of Artificial Intelligence. *Masyarakat LPPM UMJ*. https://jurnal.umj.ac.id/index.php/semnaskat

Aprinastya, R., Jazman, M., Syaifullah, S., Rahmawita, M., Siregar, S., & Saputra, E. (2024). Comparative Analysis of Naïve Bayes Classifier and Support Vector Machine for Multilingual Sentiment Analysis : Insights from Genshin Impact User Reviews. *JUSIFO: Jurnal Sistem Informasi*, *10*(2), 117–126. https://doi.org/10.19109/jusifo.v10i2.24876

Artanto, F. A. (2024). Support Vector Machine Berbasis Particle Swarm Optimization Pada Analisis Sentimen Anggota KPPS. *Jurnal Fasilkom*, *14*(1), 75–79. https://doi.org/10.37859/jf.v14i1.6795

Atmaja, F., Wahyuni, E. D., & Agussalim. (2025). Analisis sentimen berbasis aspek pada sistem layanan pengaduan masyarakat di Kota Surabaya menggunakan metode latent Dirichlet allocation dan naïve Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(1), 527–534. https://doi.org/10.36040/jati.v9i1.12438

Damayanti, E., Vitianingsih, A. V., Kacung, S., & Cahyono, D. (2024). Sentiment Analysis of Alfagift Application User Reviews Using Long Short-Term Memory ( LSTM ) and Support Vector Machine ( SVM ) Methods. *DECODE: Jurnal Pendidikan Teknologi Informasi*, *4*(2), 509–521. https://doi.org/10.51454/decode.v4i2.478

Faisal, M. R., Kartini, D., Saragih, T. H., & Arrahimi, A. R. (2022). *Belajar Data Science: Text Mining Untuk Pemula I*. https://www.researchgate.net/publication/359619425

Kacung, S., Bagyana, C. P. P., & Cahyono, D. (2024). Analisis sentimen terhadap layanan Samsat Digital Nasional (Signal) menggunakan metode SVM. *Jurnal Mnemonic*, *7*(1), 118–122. https://doi.org/10.36040/mnemonic.v7i1.9557

Komarudin, A., & Hilda, A. M. (2024). Analisis Sentimen Ulasan Aplikasi Identitas Kependudukan Digital Pada Play Store Menggunakan Metode Naïve Bayes. *Computer Science (CO-SCIENCE)*, *4*(1), 28–36. https://doi.org/10.31294/coscience.v4i1.2955

Maulida, N., Suarna, N., & Prihartono, W. (2024). Analisis Ulasan Sentimen Aplikasi Mobile Jkn Dengan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1651–1658. https://doi.org/10.36040/jati.v8i2.9105

May, I. P. A., & Fanida, E. H. (2022). Analisis efektivitas aplikasi Wargaku Surabaya dalam menunjang pelayanan publik masyarakat Kota Surabaya. *Publika*, *11(1)*, 1553–1568. https://doi.org/10.26740/publika.v11n1.p1553-1568

Motger, Q., Oriol, M., Tiessler, M., Franch, X., & Marco, J. (2025). What About Emotions? Guiding Fine-Grained Emotion Extraction from Mobile App Reviews. *2025 IEEE 33rd International Requirements Engineering Conference (RE)*. https://doi.org/10.1109/RE63999.2025.00012

Nur, D., Widiyanto K, M., & Puspitaningtyas, A. (2024). Pelayanan pengaduan masyarakat melalui aplikasi "Wargaku Surabaya" sebagai perwujudan e-governance Kota Surabaya. *Triwikrama: Jurnal Ilmu Sosial*, *4*(3), 1–10.

Purnamasari, D., Bayu, A., Desy, A., Fanka, W. A. P., Reza, A., Safrila, M., Yanda, O. N., & Hidayati, U. (2023). *Pengantar Metode Analisis Sentimen*. Gunadarma Penerbit.

Putri, D. A., & Muthia, D. A. (2024). Implementasi metode lexicon based dan support vector machine pada analisis sentimen ulasan pengguna ChatGPT. *IJCIT( Indonesian Journal on Computer and Information Technology )*, *9*(2), 80–86. https://ojs.bsi.ac.id/index.php/ijcit/article/view/23948

Riehl, K., Neunteufel, M., & Hemberg, M. (2023). Hierarchical confusion matrix for classification performance evaluation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *72*(5), 1394–1412. https://doi.org/10.1093/jrsssc/qlad057

Rifaldi, D., Fadlil, A., & Herman. (2023). Teknik preprocessing pada text mining menggunakan data tweet "mental health." *DECODE: Jurnal Pendidikan Teknologi Informasi*, *3*(2), 161–171. https://doi.org/10.51454/decode.v3i2.131

Rizaldi, S. A. R., Alam, S., & Kurniawan, I. (2023). Analisis Sentimen Pengguna Aplikasi JMO (Jamsostek Mobile) Pada Google Play Store Menggunakan Metode Naive Bayes. *STORAGE: Jurnal Ilmiah Teknik Dan Ilmu Komputer*, *2*(3), 109–117. https://doi.org/10.55123/storage.v2i3.2334

Sathyanarayanan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, *27*(4S), 4023–4031. https://doi.org/10.53555/AJBR.v27i4S.4345

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