Jurnal ICT : Information and Communication Technologies, 17 (1) (2026) 98-108

Published by: Marq & Cha Institute

Jurnal ICT : Information and Communication Technologies

Journal homepage: www.ejournal.marqchainstitute.or.id/index.php/JICT

Analysis of Public Sentiment Toward Mental Health on Social Media Using Naïve Bayes

Wiji Lestari Sitorus[1], Zuli Agustina Gultom[2]

[1,2] Program Studi Sistem Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Article Info

Article history

Received : Apr 10, 2026
Revised : Apr 27, 2026
Accepted : Apr 30, 2026

Keywords:

Mental Health ;
Sentiment Analysis;
X(Twitter);
Multinomial Naïve Bayes;
TF-IDF;
Explainability .

Abstract

Mental health is a global issue that has garnered significant public attention on social media. The platform X (formerly Twitter) is widely used by the public to openly express emotional conditions, yielding vast amounts of unstructured textual data. This research aims to analyze public sentiment regarding mental health issues on social media X using the Multinomial Naïve Bayes algorithm combined with Term Frequency-Inverse Document Frequency (TF-IDF) word weighting. The dataset consists of 9,000 tweets written in Indonesian, collected between February 15 and 27, 2025, using the keywords kesehatan_mental (mental health), stress (stress), kecemasan (anxiety), and depresi (depression). To enhance data quality, a comprehensive text preprocessing pipeline was implemented, including cleaning, case folding, word normalization (using a 59-entry mapping dictionary), tokenizing, stopword removal, and stemming. The performance of the classification model was evaluated using a confusion matrix on 1,800 test data. The results demonstrate that the Multinomial Naïve Bayes model achieved a high accuracy of 90.78% and a macro average F1-score of 90.75%. Specifically, the positive sentiment class yielded a precision of 96.22% and a recall of 84.89%, while the negative sentiment class achieved a precision of 86.48% and a recall of 96.67%. Furthermore, this study integrates the classification model into a web-based system equipped with an explainability feature that visualizes word contributions to the sentiment outcomes. This research contributes an interpretative, informative, and efficient computational approach for monitoring public sentiment trends toward mental health issues on Indonesian social media.

Corresponding Author:

Wiji Lestari Sitorus,
Program Studi Sistem Informasi,
Universitas Muhammadiyah Sumatera Utara,
Jl. Kapt. Mukhtar Basri No. 3, Glugur Darat II, Kec. Medan Timur, Kota Medan, Sumatera Utara, 20238, Indonesia
Email: wiji80601@gmail.com

This is an open access article under the CC BY-NC license.

1. Introduction

Mental health is a state of well-being that enables an individual to manage stress, work productively, and interact effectively with others. In recent years, mental health has become a global concern due to the rising number of people experiencing psychological disorders. According to a report by the World Health Organization (2022), nearly one billion people worldwide live with mental disorders.

This situation indicates that mental health is not merely an individual issue but also a social challenge requiring serious attention from various stakeholders (Mohammed & Hassani, 2025).

Advances in digital technology have transformed how society communicates its opinions and personal experiences. Social media has become one of the primary platforms people use to openly express their emotional states. The X platform allows users to share their opinions in real-time through short text posts that are easily accessible to the public. The open nature of social media generates a vast amount of textual data that can be leveraged to understand public perceptions regarding mental health issues (Fattah & Purnawansyah, 2022).

Social media data is generally unstructured and contains informal language such as slang, abbreviations, typos, and non-alphabetic symbols. These characteristics make manual analysis difficult, especially as the volume of data continues to grow. Therefore, a computational approach is needed to process text data into more meaningful and systematic information. One such approach used in processing unstructured data is data mining. Data mining enables the identification of patterns and hidden information within large datasets through the application of machine learning algorithms and data processing techniques. In the context of social media, this approach can be used to understand public opinion patterns regarding a specific issue (Dwiatmoko & Imelda, 2025).

Sentiment analysis is a component of data mining used to identify a person's opinions, emotions, or attitudes toward a topic based on digital text data. This technique can categorize public opinions into positive or negative sentiment groups, thereby providing a more objective picture of public perception. Sentiment analysis is important in the context of mental health because it can help understand public emotional trends based on social media posts (Wicaksono & Apriana, 2022).

Text data processing in sentiment analysis is generally performed using a text mining approach. Text mining is a technique for extracting important information from unstructured textual data so that it can be transformed into more organized and easily analyzable data. This process involves various stages of text processing to improve data quality prior to classification (Mailo et al., 2020)

The preprocessing stage is a crucial part of text mining as it serves to clean and standardize text data. The cleaning process is performed to remove symbols, numbers, URLs, hashtags, and other characters irrelevant to the classification process. This stage helps reduce noise in the dataset, thereby improving data quality for processing by classification algorithms. In addition to cleaning, preprocessing also involves case folding to standardize all letters to lowercase to prevent differences in word interpretation due to the use of capital letters. After that, word normalization is performed to convert non-standard words, slang, or typos into standard word forms according to Indonesian language rules (Fattah & Purnawansyah, 2022).

The tokenization step is performed to break sentences down into tokens or individual word segments so that the algorithm can perform feature extraction more effectively. This is followed by stopword removal to eliminate common words that do not significantly contribute to determining the text's sentiment. The preprocessing process concludes with stemming to convert inflected words into their base forms, ensuring that morphological variations of a word are recognized as the same entity (Syahputra & Kurniawan, 2024).

Sentiment analysis regarding mental health issues on social media has been extensively explored using various classification methods. One approach proven to be effective is the use of the Naïve Bayes algorithm on Twitter data related to this topic. Through the application of this method, the system was able to group public sentiment very well and achieved an accuracy rate of 89%. These results confirm that this probability-based algorithm performs reliably in recognizing and classifying textual data related to mental health issues (Aulia et al., 2020).

The application of the Naïve Bayes Classifier method combined with TF-IDF word weighting has been used to detect indications of suicidal tendencies among college students via the social media platform Twitter. This approach has proven to yield reasonably sensitive results, with an accuracy rate of 90.24%. Nevertheless, the resulting classification model still has limitations regarding generalization because the amount of data used in the testing process tends to be limited, so its performance is not yet optimal when applied on a larger scale (Ainnur Rafli, 2024).

Sentiment analysis regarding specific psychological disorders, such as bipolar disorder, has also been conducted on Twitter using the Naïve Bayes algorithm. The application of this method proved highly effective for text-based classification, yielding a high accuracy rate of 92.11%. Although it demonstrated highly reliable performance in identifying relevant text characteristics, the focus of this study remained strictly on bipolar disorder and has not yet been tested to encompass or describe general mental health conditions (Sativa & Silaen, 2022).

The Naïve Bayes algorithm has also been applied to analyze sentiment related to cyberbullying on the X platform. This method has proven capable of effectively distinguishing between negative and positive comments, achieving an accuracy rate of 86%. In addition to delivering optimal results, the implementation of this system also demonstrated high effectiveness when dealing with the characteristics of Indonesian-language text data that uses non-standard or informal language (Arfan et al., 2024).

Another approach to mental health sentiment analysis on Twitter was conducted using the K-Nearest Neighbor (KNN) algorithm. The application of this method is considered quite capable and able to cluster user sentiments effectively. However, this distance-based algorithm has a significant weakness in terms of scalability, especially when forced to process large-scale datasets with very high feature dimensions (Wicaksono & Apriana, 2022).

Based on previous studies, it can be concluded that mental health sentiment analysis still faces challenges in processing unstructured text data and in the computational efficiency of classification models. Some studies employ highly complex methods that require significant computational resources, while others are limited by the size of their datasets or the scope of mental health issues analyzed.

Therefore, this study uses the Multinomial Naïve Bayes algorithm with TF-IDF weighting to analyze public sentiment toward mental health issues on social media platform X. This study implements a comprehensive preprocessing stage that includes cleaning, case folding, word normalization, tokenization, stopword removal, and stemming to improve data quality before the classification process is carried out.

The contribution of this study lies in the application of a website-based mental health sentiment classification model using a large dataset of Indonesian-language social media posts on X. Additionally, this study implements explainability features to highlight the contribution of individual words to sentiment classification results, making the analysis process more interpretable and informative compared to previous studies

2.Research Methodology

The research stages to be conducted in this study are as follows:

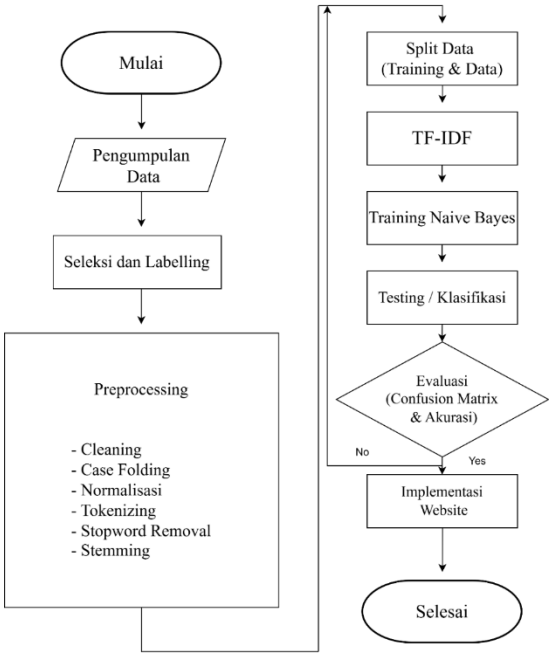

Figure 1. Research Stages

1. Data Collection

The first stage of this study is the collection of secondary data in the form of unstructured text derived from tweets by users of social media platform X regarding mental health issues. Data were collected during the period of February 15–27, 2025, using the keywords mental_health, stress, anxiety, and depression. The total dataset obtained consisted of 9,000 tweets, comprising 6,848 tweets (76.09%) on mental_health, 1,721 tweets (19.12%) on stress, 425 tweets (4.72%) on anxiety, and 6 tweets (0.07%) on depression. This distribution indicates the dominance of discussions on mental health in general on X social media during the data collection period.

2. Data Selection and Labeling

The second stage involves data selection and labeling. The data selection stage ensures that only tweets relevant to mental health issues are used in the study. Tweets that are duplicates, lack text content, or are unrelated to the research topic are then removed before the labeling and preprocessing processes are carried out. The data labeling process is performed manually by reading each tweet collected from the social media platform X to determine sentiment categories based on the meaning and context of the sentences. Tweets are categorized as positive sentiment if they contain support, positive views, or positive experiences related to mental health, while negative sentiment is assigned to tweets containing complaints, emotional distress, stress, anxiety, or depression.

3. Data Preprocessing

The third stage is data preprocessing, which includes cleaning, case folding, normalization, tokenization, stopword removal, and stemming. The first step in the data preprocessing pipeline is:

Cleaning is the first step in the preprocessing pipeline, aimed at removing text components that are noise and do not contribute to sentiment analysis.

Table 1 Cleaning Process

No	Before Cleaning	After Cleaning
1	@user Burnout parah banget! #kesehatan https://t.co/abc	Burnout parah banget Kesehatan
2	RT @dokter: Stress level 100 bgt nih https://bit.ly/xyz	Stress level bgt nih

3	Semangat! Jaga kesehatan mental ya guys	Semangat Jaga kesehatan mental ya guys

Case folding standardizes all characters in the text to lowercase, so that words with different capitalization are treated identically by the model; for example, "Burnout" and "burnout" become the same token. This process prevents unnecessary feature duplication.

Table 2: Example of the Case Folding Process

No	Before Case Folding	After Case Folding
1	Burnout parah banget kesehatan	burnout parah banget kesehatan
2	Stress level bgt ini	stress level bgt ini
3	Semangat Jaga kesehatan mental ya guys	semangat jaga kesehatan mental ya guys

Word normalization converts non-standard words, abbreviations, and slang into standard forms using a mapping dictionary consisting of 59 entries. This dictionary includes common Indonesian social media abbreviations and English mental health terms frequently used in informal contexts, such as burnout → kelelahan and overthinking → berpikir.

Table 3: Examples of the Word Normalization Process

No	Original Word	Normal Word	Explanation
1	Gak	Tidak	Singkatan negasi
2	Sudah	Sudah	Singkatan
3	Burnout	Kelelahan	Istilah Inggris
4	Overthinking	Berpikir	Istilah Inggris
5	Stress	Stress	Ejaan baku
6	Anxiety	Cemas	Istilah Inggris
7	Bgt	Banget	Singkatan

Tokenization breaks down normalized sentences into individual word units (tokens) by separating them based on spaces. These tokens form the basis for building the model's vocabulary and calculating TF-IDF weights in the next stage.

Table 4 Example of the Tokenization Process

No.	Normalized Text	Token Result
1	kelelahan parah banget Kesehatan	[kelelahan, parah, banget, kesehatan]
2	stres level banget ini	[stres, level, banget, nih]

Stopword removal removes common words that carry no semantic weight for sentiment (e.g., "the," "in," "that," and "and"). This reduces the feature dimension and highlights relevant keywords.

Table 5: Example of the Stopword Removal Process

No.	Before Stopword Removal	After Stopword Removal
1	[kelelahan, parah, banget, kesehatan]	[kelelahan, parah, kesehatan]
2	[stres, level, banget, nih]	[stres, level]

3	[semangat, jaga, kesehatan, mental, ya, guys]	jaga,	[semangat, kesehatan, mental]

Stemming reduces inflected words to their root form by removing prefixes and suffixes. Stemming is performed by first checking for suffixes, then prefixes, provided that the remaining word is at least 3 characters long.

Table 6: Examples of the Stemming Process

No.	Real Token	Stem Result	Deleted
1	Menjaga	Jaga	me-
2	Kelelahan	Lelah	ke- + -an
3	Diperhatikan	perhatikan	di-
4	Kesehatan	Sehat	ke- + -an
5	Memikirkan	Pikir	me- + -kan

4. Data Split

The fourth step is data splitting. Once the preprocessing is complete, the dataset is divided into two parts: training data and testing data. The training data is used to build the classification model, while the testing data is used to evaluate the model's performance.

5. TF-IDF Weighting

The fifth step is TF-IDF weighting, where the text is converted into a numerical representation using the Term Frequency–Inverse Document Frequency (TF-IDF) method, a word weighting technique that assesses the importance of a word within a document relative to the entire dataset.

6. Naïve Bayes Classification

The sixth step involves using the text represented in TF-IDF format as the basis for training the Naïve Bayes algorithm, where the model learns the patterns of word occurrence across each sentiment category. The trained model is then used to classify text into positive or negative sentiment categories based on the highest probability generated by the Naïve Bayes calculation. Mathematically, Bayes' theorem, which forms the basis of this algorithm, can be formulated as follows:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (1)$$

Description:

X is data of an unknown class (evidence).

H is the hypothesis that data X belongs to a specific class.

P(H|X) is the probability of hypothesis H given data X (posterior probability).

P(H) is the prior probability of hypothesis H.

P(X|H) is the probability of data X occurring assuming hypothesis H is true (likelihood).

In the context of mental health sentiment analysis, this algorithm calculates the probability of words appearing in a document to determine whether a text falls into the positive or negative sentiment category.

7. Evaluating the Confusion Matrix

The seventh step involves evaluating the classification results using a confusion matrix to determine the model's accuracy. The following are the formulas for the four evaluation metrics used in the confusion matrix:

Accuracy is used to measure the model's precision in predicting the data as a whole

Accuracy = (TP + TN) / (TP + TN + FP + FN) (2)

Precision is used to measure a model's accuracy in predicting positive classes.

Precision = TP / (TP + FP) (3)

Recall is used to measure a model's ability to identify all positive data.

Recall = TP / (TP + FN) (4)

The F1-score is the harmonic mean of precision and recall, used to balance these two metrics.

F1 = (2 × Precision × Recall) / (Precision + Recall) (5)

8. Website Implementation

The final step involves integrating the developed model into a web-based system so that users can enter text and obtain sentiment analysis results directly through the available interface.

3.Results and Discussion

1. Home Page Display

This study uses Indonesian-language tweet data obtained from the social media platform X regarding mental health issues. A total of 9,000 tweets were collected using the keywords kesehatan_mental, stress, kecemasan, and depresi. The home page is the first page that appears when a user accesses the system. This page displays real-time system statistics: the total number of predictions made, the percentage of positive and negative sentiment, the accuracy of the active model, and the number of vocabulary terms.

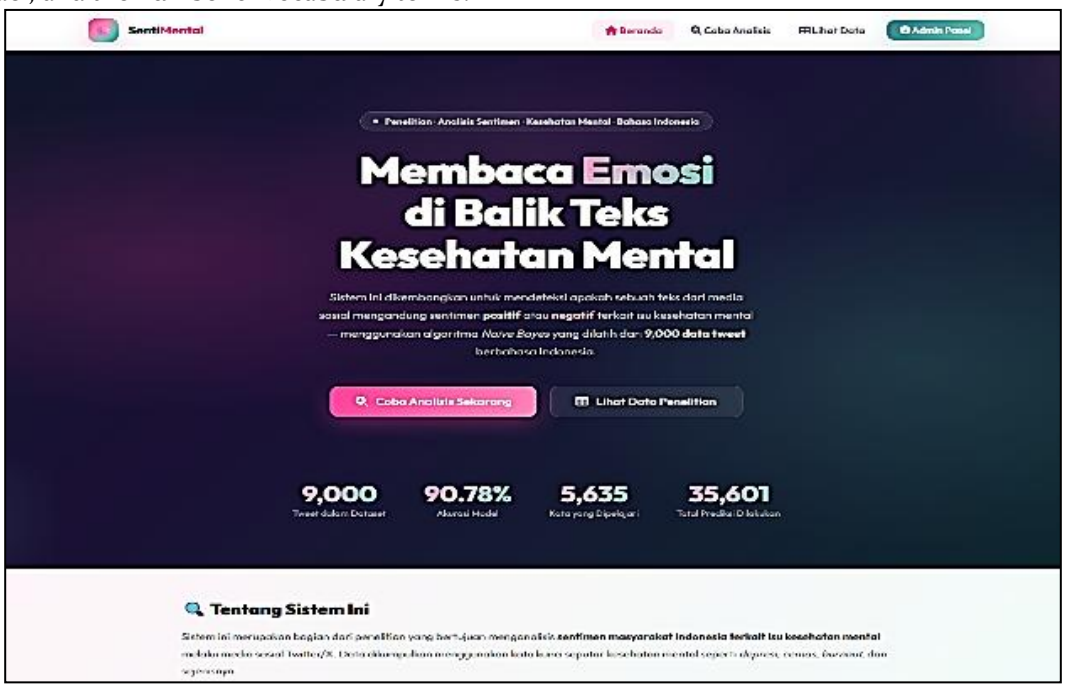

Figure 2. Home Page Display

3. Sentiment Analysis Page Display

Figure 3 shows the Sentiment Analysis Page, which displays the interface of a web-based mental health sentiment analysis system using the Multinomial Naïve Bayes algorithm. The system is used to automatically analyze the sentiment of user text or emotional outpourings related to mental health. In the test example, the user entered the sentence “My depression is getting worse, nobody cares, I feel so lonely and hopeless” into the text input field.

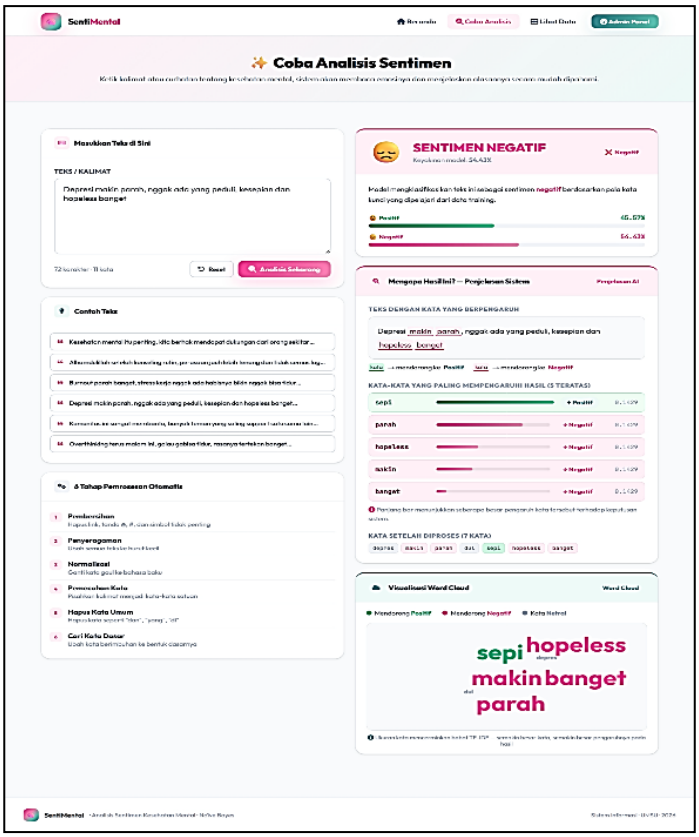

Figure 3. Sentiment Analysis Page Display

The classification results show that the system categorized the text as negative sentiment with a confidence level of 54.43%, while the probability of positive sentiment was 45.57%. The classification results were determined based on word patterns learned by the model from the training data using the TF-IDF method and the Multinomial Naïve Bayes algorithm. Before the classification process was carried out, the system applied preprocessing steps that included cleaning, case folding, normalization, tokenization, stopword removal, and stemming to improve the quality of the text data.

This research also implements an explainability feature to display the contribution of words to sentiment classification results. This feature allows the system to provide an interpretation of the words that most influence the determination of a tweet's sentiment category. With this feature, the classification results become easier to understand and more informative for users.

3. Admin Login Page (login.php)

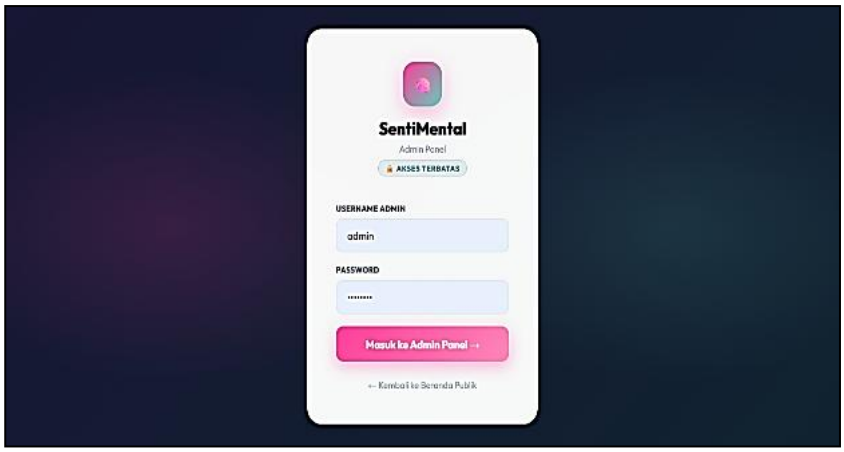

Figure 4. Admin Login Page

The login page serves as the entry point to the administrator area. The system uses PHP session mechanisms to maintain authentication status. If authentication is successful, the admin is redirected to the dashboard; if it fails, an error message is displayed. Direct access to the admin URL without an active session will be automatically redirected to the login page via the requireAdmin() function in config.php.

4. Admin Dashboard Page

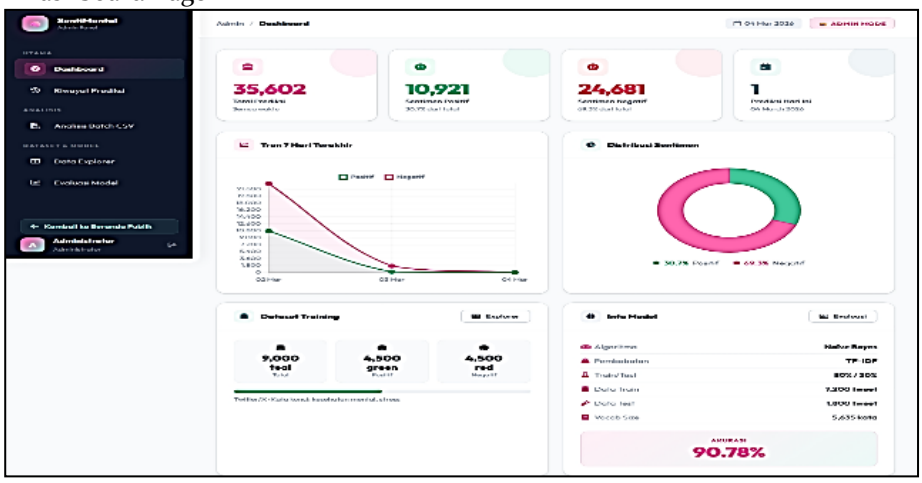

Figure 5. Admin Dashboard Page

The dashboard page serves as the system's monitoring hub and is accessible only to administrators. This page displays four real-time statistics cards (total predictions, number of positives, number of negatives, today's predictions), a line graph showing trends over the past seven days, a pie chart of sentiment distribution, a table of the ten most recent predictions, and a navigation sidebar with information on the active model configuration.

5. Model Evaluation Page

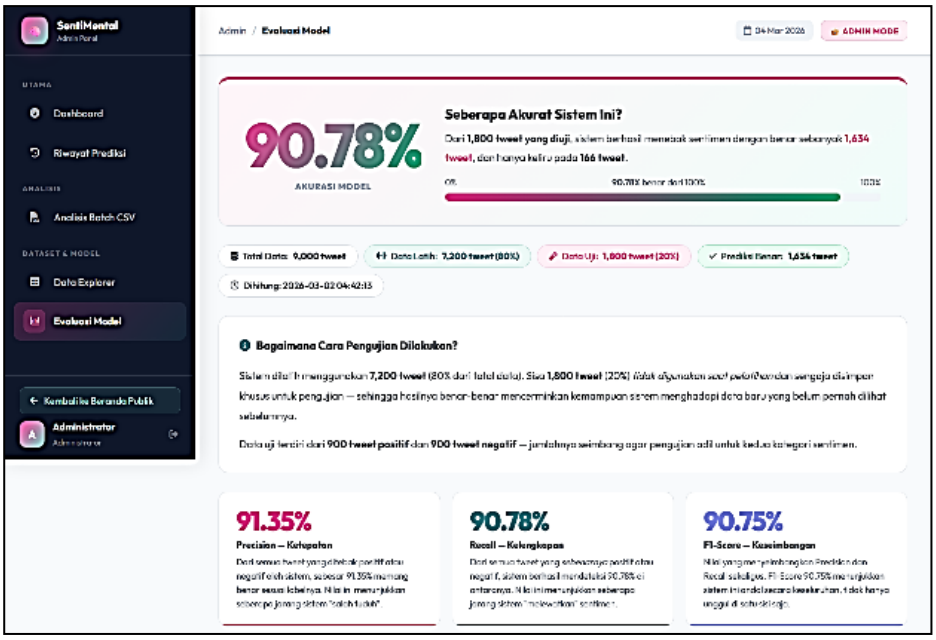

Figure 6. Model Evaluation Page Display

In Figure 6, the TF-IDF-weighted Naïve Bayes Multinomial Model achieved an accuracy of 90.78% on 1,800 test data points. Of the total data, the model correctly classified 1,634 tweets, while the remaining 166 tweets (9.22%) were misclassified. Additionally, the model produced a macro average F1-Score of 90.75%, indicating good classification performance for both sentiment classes. For the positive class, the model achieved a precision of 96.22%, a recall of 84.89%, and an F1-Score of 90.20%. Meanwhile, for the negative class, the model achieved a precision of 86.48%, a recall of 96.67%, and an F1-Score of 91.29%.

4.Conclusion

This study successfully applied the Multinomial Naïve Bayes algorithm to analyze sentiment regarding mental health issues on the social media platform X through preprocessing and TF-IDF weighting. The test results showed that the model achieved an accuracy of 90.78% with a macro average F1-score of 90.75%, demonstrating its ability to effectively classify positive and negative sentiment. Additionally, the system is equipped with an explainability feature to display the words that influence the classification results. This study contributes to the development of text mining-based sentiment analysis for mental health issues on social media. The research findings are expected to serve as a reference for developing efficient sentiment classification systems, particularly for unstructured Indonesian-language text data. Future research is recommended to use a larger and more balanced dataset to optimize classification results. Furthermore, subsequent studies could include additional sentiment categories, such as neutral, and compare the performance of the Naïve Bayes algorithm with other classification methods—such as Support Vector Machine, Random Forest, or Deep Learning—to achieve more accurate results.

References

Ainnur Rafli, & K. (2024). Indonesian Journal of Computer Science. 13(1), 1038–1049.

Arfan, I. S., Fauziah, S., & Nawangsih, I. (2024). Sentiment Analyst of Cyber Bullying in X Using Naïve Bayes Algorithm Analisa Sentimen Terhadap Cyber Bullying di X Menggunakan Algoritma Naïve Bayes. 4(October), 1411–1419.

Aulia, K., Amelia, L., & Mental, K. (2020). ANALISIS SENTIMEN TWITTER PADA ISU MENTAL HEALTH DENGAN ALGORITMA KLASIFIKASI NAIVE BAYES. 6(2), 60–65.

Dwiatmoko & Imelda, 2025. (2025). ANALISIS SENTIMEN KESEHATAN MENTAL DI TWITTER

MENGGUNAKAN NAIVE BAYES CLASSIFIER DAN K- NEAREST NEIGHBOR. 6(1), 1–9.
Fattah & Purnawansyah, 2022. (2022). Analisis sentimen terhadap body shaming pada twitter menggunakan metode Naïve Bayes Classifier. 3(2), 61–71.
Mailo, F. F., Lazuardi, L., Manajemen, D., Fakultas, K., Masyarakat, K., & Mada, U. G. (2020). Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia. 4(1).
Mohammed, I., & Hassani, H. (2025). Mining Mental Health Signals: A Comparative Study of Four Machine Learning Methods for Depression Detection from Social Media Posts in Sorani Kurdish. 1–13.
Sativa, O., & Silaen, D. (2022). Analisis Sentimen Mengenai Gangguan Bipolar Pada Twitter Menggunakan Algoritma Naïve Bayes. 6(2), 63–73.
Syahputra, P., & Kurniawan, R. (2024). Analisis Sentimen Terhadap Kesehatan Mental Remaja Menggunakan Metode Naive Bayes. 5(4), 1216–1224. https://doi.org/10.47065/josh.v5i4.5644
Wicaksono, M. L., & Apriana, D. (2022). ANALISIS SENTIMEN KESEHATAN MENTAL MENGGUNAKAN K-NEAREST NEIGHBORS PADA SOSIAL MEDIA TWITTER SENTIMENT ANALYSIS OF MENTAL HEALTH USING K-NEAREST NEIGHBORS ON SOCIAL MEDIA TWITTER. 19(2), 98–103.