Emotion Detection on Platform X Comment with Naive Bayes Classification

Fulzan Abid[1], Muchamad Kurniawan[2]*, Hamdan Bahalwan[3], Andy Rachman[4], Faza Wahmuda[5], Syahri Muharom[6], Anwar Sodik[7]
[1,2,4]Informatics Department, Institut Teknologi Adhi Tama Surabaya, Indonesia
[3,5]Product Design Department, Institut Teknologi Adhi Tama Surabaya, Indonesia.
[6]Electrical Engineering Department, Institut Teknologi Adhi Tama Surabaya, Indonesia
[7]Information System Department, Institut Teknologi Adhi Tama Surabaya, Indonesia.
[1]dzulfanabid@gmail.com; [2]muchamad.kurniawan@itats.ac.id*; [3]hamdan.despro@itats.ac.id; [4]andy.rach1910@itats.ac.id;
[5]faza.despro@itats.ac.id; [6]syahrimuharom@itats.ac.id; [7]anwar@itats.ac.id
*corresponding author

ABSTRACT

This study aims to develop an effective emotion-detection model for Indonesian-language Twitter comments using a lightweight, interpretable machine learning approach. The proposed method combines the Naive Bayes Classifier (NBC) with Term Frequency–Inverse Document Frequency (TF–IDF) for text feature extraction. The dataset used in this study comprises 3,115 Indonesian-language comments from the publicly available X Emotion Dataset. Emotion detection on Platform X is essential given the platform's high activity and the need for automated monitoring of public sentiment and online behaviour. Four data split scenarios, among them 60:40, 70:30, 80:20, and 90:10, were evaluated to measure the model's accuracy, recall, and precision in classifying emotions into anger, happiness, and sadness. The experimental results show that the 80:20 ratio achieved the highest accuracy of 68.86%, providing an optimal balance between learning efficiency and generalization capability. The *anger* class consistently achieved the highest recognition rate, while the happy and sad classes showed moderate results due to overlapping linguistic characteristics. Although this study is limited to three emotion classes and a single algorithm, the findings demonstrate that the Naive Bayes–TF–IDF combination remains robust for emotion classification in resource-limited languages. This research contributes an interpretable, computationally efficient framework for social media sentiment analysis and digital behavioural studies in the Indonesian language context.

Keywords: Emotion Detection; Naive Bayes Classifier; TF-IDF; X-platform; Natural language processing.

Article History		
Received : October, 27th 2025	Accepted: December, 1st 2025	Published: December, 5th 2025

I. INTRODUCTION

In the digital era, social media platforms have become a primary medium for users to share thoughts, opinions, and emotions in real time. The widespread use of these platforms has generated massive amounts of unstructured textual data that contain valuable insights into public sentiment and emotional expression [1]. According to DataReportal (2023) [2], Indonesia ranks among the top countries for social media usage, with millions of users actively engaging on X (formerly Twitter). These users frequently express diverse emotions, including happiness, anger, sadness, and even hate speech, especially in discussions about politics, entertainment, and public issues [3][4].

Emotions play a significant role in influencing user interactions and online behaviour. According to [5], there are six basic emotions universally recognized across cultures: happiness, sadness, anger, fear, disgust, and surprise. However, in textual communication, especially on microblogging platforms like X, emotional expression is often limited to textual cues, such as word choice, sentence structure, and emotive vocabulary, rather than facial expressions or tone. This makes text-based emotion detection a challenging yet essential task for understanding public mood, opinion dynamics, and digital discourse patterns [6].

The rise of text mining and machine learning techniques has significantly improved the accuracy of emotion and sentiment classification [7]. Emotion detection, as a subfield of natural language processing (NLP), aims to classify textual content into emotional categories by analyzing linguistic features and patterns. Several approaches have been proposed in the literature, including lexicon-based, machine learning-based, and hybrid models [8]. Lexicon-based approaches rely on predefined emotion dictionaries, such as EmoLex or the APA Emotion Dictionary, in which each word is associated with an emotional label. However, these methods are limited in their ability to handle context, sarcasm, and word ambiguity.

Machine learning methods, on the other hand, utilize statistical and probabilistic techniques to infer emotional categories from annotated datasets. Various algorithms have been applied in this field, including Support Vector Machines (SVM), Logistic Regression, Random Forests, and Naive Bayes Classifiers (NBC) [9][10][11]. More recently, deep learning models such as neural

networks have achieved superior results in multi-class emotion recognition [8]. Nevertheless, traditional models such as Naive Bayes remain widely used due to their simplicity, interpretability, and computational efficiency, especially for medium-sized datasets and multilingual corpora [12][13].

In comparative studies, Naive Bayes has demonstrated robust performance in emotion and sentiment classification. For instance, Machová et al. (2023) [8] reported that Naive Bayes achieved an average performance score of 0.56, outperforming the Lexicon-based method (0.26) and SVM (0.47). Similarly, Azmin (2019) found that Naive Bayes reached 78.6% accuracy in classifying emotions from Bangla text, compared to 71.6% for SVM. Another study on Twitter data for depression detection found that Naive Bayes achieved 89% accuracy, 90% precision, 89% recall, and 90% F1-score, slightly surpassing Random Forest [13]. These results suggest that Naive Bayes provides a competitive balance between accuracy, computational cost, and generalization ability.

In addition to classifier choice, the feature extraction method also plays a critical role in improving classification accuracy. The Term Frequency–Inverse Document Frequency (TF-IDF) technique has been widely adopted as an effective method for text representation in NLP tasks [14][15]. TF-IDF assigns higher weights to rare yet informative words while downweighting common ones, enabling the model to focus on distinctive emotional indicators [16]. Studies comparing TF-IDF with Bag of Words (BoW) have shown that TF–IDF consistently achieves better accuracy and interpretability [17][18].

Several research works have explored emotion detection in Indonesian social media contexts. [19] Conducted sentiment analysis on global Twitter reactions to the city of Medan using Naive Bayes and cross-validation, while [20] applied the same algorithm to evaluate employee performance in a corporate dataset. [3] Specifically analyzed emotions toward presidential candidates in Indonesia's 2024 general election using Naive Bayes, achieving promising classification results. Meanwhile, [21] examined both emotion and hate-speech classification in Indonesian tweets, demonstrating that text-based models can effectively detect psychological tone and aggression patterns online.

Evaluation of machine learning models typically involves statistical metrics that measure their predictive performance. Commonly used metrics include Accuracy, Precision, and Recall, which are derived from the confusion matrix [22][23]. Accuracy measures the proportion of correctly predicted labels, precision assesses the relevance of positive predictions, and recall determines how many actual positives were correctly identified. These metrics collectively provide a comprehensive evaluation of the model's performance, particularly on imbalanced datasets, which are typical in emotion classification [6].

While several studies have successfully applied machine learning to emotion detection in different languages, such as Arabic [11], Bangla [12], and English [8], research on Indonesian-language emotion detection remains limited. The morphological richness, informal word usage, and cultural variation in emotional expression present unique challenges for computational analysis. Hence, there is a strong need for a lightweight yet accurate model that can process Indonesian text efficiently and produce reliable emotion classification results.

Naive Bayes is chosen in this study not only for its interpretability but also for its computational efficiency, which is particularly suitable for medium-sized Indonesian datasets. Prior studies have shown that Naive Bayes requires significantly fewer computational resources than SVMs or deep learning models, making it ideal for practical applications and baseline modelling. Although several studies have applied Naive Bayes or other machine learning models to Indonesian sentiment and emotion analysis, most evaluate model performance using only a single train–test split. Consequently, the effect of different data proportions on model stability and generalization remains underexplored. This gap is particularly relevant for Indonesian-language NLP, where datasets are often limited and imbalanced.

Based on these considerations, this study aims to develop an emotion detection model for Indonesian-language Twitter comments using the Naive Bayes Classifier with TF–IDF feature extraction. The model focuses on three primary emotions: happy, sad, and angry, which are the most frequently expressed in online discourse [24]. The research involves data pre-processing, TF–IDF feature weighting, Naive Bayes model training, and performance evaluation using Accuracy, Precision, and Recall metrics. This study contributes by: (1) evaluating the performance of a lightweight and interpretable Naive Bayes–TF–IDF model across four train–test scenarios, (2) analyzing robustness under class imbalance, and (3) providing empirical insights for practical implementation in low-resource Indonesian social media emotion detection. While Naive Bayes and TF–IDF are well-established methods, their performance consistency under multiple data-splitting scenarios for Indonesian emotion detection has not been extensively explored. This study contributes by analyzing model robustness across varying proportions of training data, providing insights for low-resource and real-time analytic environments.

II. METHOD

Figure 1 shows the research methodology for emotion detection in X-platform comments using Naive Bayes (NB) classification, illustrating the entire research process from data collection to model evaluation. The process begins with crawling data from Platform X to obtain user comments, which serve as the primary source for analysis. These comments are then pre-processed, including case folding, tokenization, stopword removal, and stemming, to clean the text data and prepare it for processing. Model evaluation was conducted using several performance metrics, including accuracy, precision, recall, and F1-score, to assess the model's ability to classify emotions consistently and accurately.

DOI: http://dx.doi.org/10.25139/ijair.v7i1.11156

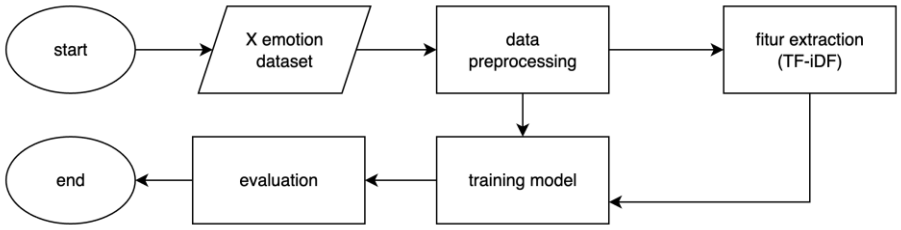


Fig. 1. Research Methodology

A. *Data Description*

In this study, the original dataset was obtained from the *Emotion Dataset from Indonesian Public Opinion*, an open-source Twitter-based emotion corpus published on GitHub (https://github.com/Ricco48/Emotion-Dataset-from-Indonesian-Public-Opinion). The dataset contains 7,080 Indonesian tweets, labelled into six emotion categories: anger (1,130), fear (911), joy (1,275), love (760), sadness (1,003), and neutral (2,001).

From these original six classes, the present study focuses only on three primary emotions: anger, joy, and sadness. This selection is grounded in both empirical analysis and theoretical considerations. Our preliminary examination of conversational text in Indonesian online discourse indicates that these three emotions are the most consistently and clearly expressed in short-form text. In contrast, emotions such as love and fear tend to be highly context-dependent and are often communicated through implicit or ambiguous linguistic cues, which reduce classification reliability. This observation also aligns with prior affective computing literature, which recognizes anger, joy/happiness, and sadness as core, universally distinguishable emotional primitives commonly used in text-based emotion recognition benchmarks.

Therefore, restricting the dataset to these three emotions improves signal clarity and allows the model to focus on emotion categories that are both frequently occurring and reliably represented in textual form. The filtered dataset used in this research consists exclusively of comments from these three classes, serving as the final training and testing data for model development.

B. *Data Pre-processing*

Raw social media data typically contains noise, such as emojis, links, mentions, repeated punctuation, and inconsistent capitalization. To ensure the dataset is suitable for analysis, pre-processing is essential to convert unstructured text into a clean, structured form [6]. In this study, data pre-processing was performed through several sequential steps:

1) Case Folding: All characters are converted into lowercase to maintain text consistency and avoid treating words like "Happy" and "happy" as different tokens.

2) Punctuation and Number Removal: All punctuation marks, numbers, and special characters that do not contribute to emotional meaning are removed.

3) Tokenization: Each sentence is split into individual word units (tokens) using a tokenizer function.

4) Stop-word Removal: Common functional words (such as "and", "the", "*yang*", "*di*") are removed since they rarely carry emotional significance [7][24].

5) Stemming: Each token is reduced to its base or root form. Since the training dataset consists of English text, English pre-processing rules are applied. However, the validation dataset contains Indonesian comments; therefore, an Indonesian stemmer (Sastrawi) is also used to ensure optimal normalization across both languages. This dual-language pre-processing approach was necessary but was not clearly articulated in the earlier version of the manuscript.

This pre-processing pipeline ensures that only meaningful, linguistically normalized words remain across both the English and Indonesian datasets, thereby facilitating accurate feature extraction and emotion classification. This pre-processing pipeline ensures that only meaningful words remain, facilitating accurate feature extraction and classification.

C. *TF-IDF*

After pre-processing, the clean text is converted into a numerical representation that machine learning models can understand. In this research, the Term Frequency–Inverse Document Frequency (TF–IDF) method is employed, as it effectively captures the significance of a word relative to a document and the corpus as a whole [14].

The Term Frequency (TF) measures how frequently a term appears within a single document, as expressed in Equation (1). Where $f_{t,d}$ denotes the frequency of term *t* in document *d*.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (1)$$

Meanwhile, the Inverse Document Frequency (IDF) measures the importance of a term across all documents, reducing the influence of terms that appear too frequently, as shown in Equation (2). Where *N* is the total number of documents, and df_t is the number of documents containing term *t*.

$$IDF(t) = \log \frac{N}{df_t} \tag{2}$$

The combined TF–IDF weight for each term is then computed using Equation (2). As demonstrated by [16], TF–IDF provides superior results to the Bag-of-Words (BoW) approach by emphasizing unique, information-rich terms and downweighting common or irrelevant ones. In this study, the TF–IDF vectorizer transforms each comment into a weighted numerical vector, which serves as the classifier's input.

$$w_{t,d} = TF(t,d) \times IDF(t) \tag{3}$$

D. Naïve Bayes Classifier

For classification, this research employs the Naive Bayes Classifier (NBC), a probabilistic model grounded in Bayes' Theorem using Equation (4). The algorithm predicts the most probable emotion class for a given text by assuming conditional independence between features (words). Despite this naive assumption, the method has consistently achieved strong performance in text classification tasks [25].

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \tag{4}$$

These components form the basis of the Bayesian approach used for text classification. The term *P(C|X)* represents the posterior probability, which is the probability that a text *X* belongs to a specific class *C* after considering the observed features. The value *P(X|C)* refers to the likelihood, indicating how probable it is to observe the text *X* if it truly comes from class *C*. Meanwhile, *P(C)* is known as the prior probability, describing how frequently class *C* appears in the dataset before any evidence is considered. The *P(X)* represents the evidence or marginal probability of the text, which acts as a normalization factor to ensure that the posterior probabilities across all classes sum to one.

In this study, the Multinomial Naive Bayes variant is used, as it is suitable for discrete features such as word frequencies or TF–IDF values. The model computes the probability for each emotion class *Ci* and selects the class with the highest posterior probability as the final label, as shown in Equation 5, where x_j represents the j^{th} term in the document vector.

$$C^* = \arg\max_{C_i} P(C_i) \prod_{j=1}^{n} P(x_j|C_i) \tag{5}$$

E. Evaluation Metrics

To assess the model's performance, three key metrics are employed: Accuracy, Precision, and Recall. These metrics are computed based on the Confusion Matrix, which records the number of correctly and incorrectly classified instances across categories [22][23]. Accuracy measures the model's overall correctness using Equation (6). Precision is the proportion of correctly predicted positive instances among all predicted positives, as defined in Equation (7). Recall (or Sensitivity) measures the proportion of correctly predicted positives among all actual positives, computed using Equation (8).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. High precision and recall values indicate that the model can accurately identify emotional content while minimizing both false alarms and missed detections. These metrics were selected because they provide a balanced view of model performance, especially for datasets with uneven class distributions, common in emotion detection tasks where some emotions appear more frequently than others [8].

III. RESULT AND DISCUSSION

Naive Bayes is chosen in this study not only for its interpretability but also for its computational efficiency, which is particularly suitable for medium-sized Indonesian datasets. Prior studies have shown that Naive Bayes requires significantly fewer computational resources than SVMs or deep learning models, making it ideal for practical applications and baseline modelling. NBC produces all results in this section. In this study, a total of 3,115 comments were used as the dataset, obtained from the X Emotion Dataset available on GitHub. The dataset was split into training and test sets across several test scenarios to evaluate the model's performance. In the first experiment, the data were split at a 60:40 ratio, with 1,869 randomly selected comments in the

training set and 1,246 in the test set. In the second experiment, a 70:30 split was used, with 2,180 comments for training and 935 for testing. The third experiment used an 80:20 split, comprising 2,492 comments in the training set and 623 in the test set. Finally, in the fourth experiment, the data were split at a 90:10 ratio, with 2,803 comments assigned to the training set and 312 to the testing set.

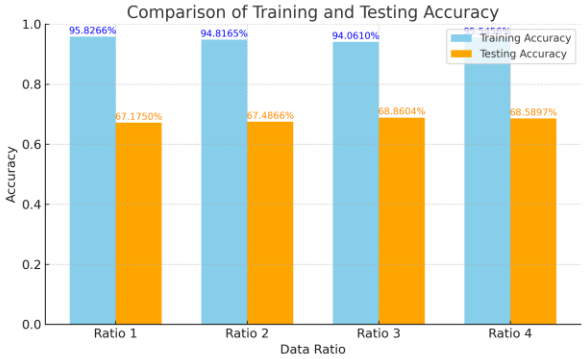
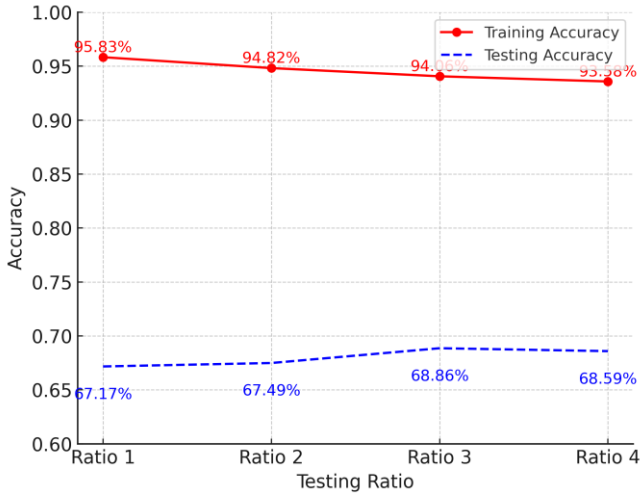


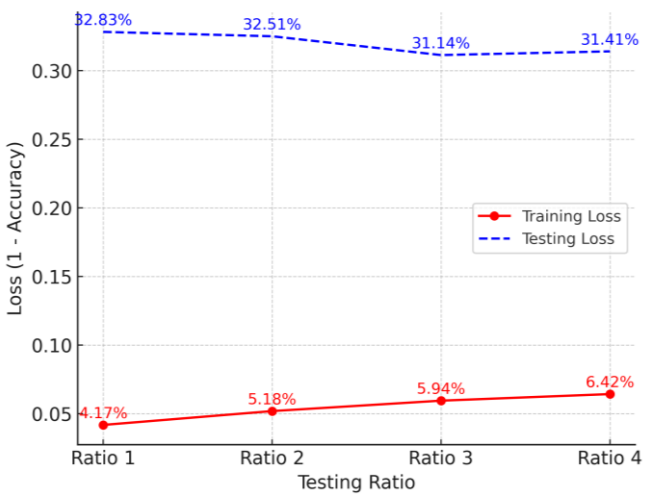
Fig.2. Comparison of Training and Testing Results

Fig. 2 compares training and test accuracy across four experimental ratios: 60:40 (Ratio 1), 70:30 (Ratio 2), 80:20 (Ratio 3), and 90:10 (Ratio 4). The graph shows that the training accuracy remains consistently high across all experiments, ranging from 94.06% to 95.83%, while the testing accuracy varies from 67.17% to 68.86%. The high training accuracy indicates that the Naive Bayes Classifier (NBC) effectively learned patterns from the training data. However, the lower testing accuracy relative to training accuracy suggests mild overfitting, where the model performs slightly better on known data than on unseen data. This is a common phenomenon in text classification tasks involving small or medium-sized datasets, especially when the dataset contains overlapping emotional vocabulary between classes, such as *happy*, *sad*, and *angry*. The highest testing accuracy, 68.86%, was achieved in the 80:20 data ratio (Ratio 3). This result indicates that the model performed optimally when trained on 80% of the data and tested on 20%. This ratio provides a good balance between model learning capacity and generalization ability, as the model has sufficient training samples to recognize emotional features while being evaluated on a substantial amount of unseen data.

In contrast, when the ratio increased to 90:10 (Ratio 4), the improvement in testing accuracy was marginal (68.59%), despite a higher amount of training data. This shows that increasing the training portion beyond a certain point does not significantly improve model performance, as the testing set becomes too small to provide a robust evaluation. Meanwhile, the 60:40 and 70:30 ratios yielded testing accuracies of 67.18% and 67.49%, respectively. These values were slightly lower than in the 80:20 scenario, indicating that a smaller training set may limit the model's ability to learn sufficient emotional patterns, resulting in slightly reduced performance.



(a) Changes in Accuracy on Training and Testing with Different Ratios



(b) Changes in Loss on Training and Testing with Different Ratios

Fig.3. Accuracy and Loss Changes Across Different Train/Test Ratios

Fig.3 illustrates the variation in model performance across four training and testing ratios (60:40, 70:30, 80:20, and 90:10) using two comparative plots: (a) changes in accuracy, and (b) changes in loss (1 – accuracy). Both subfigures examine how the balance between training and test data affects the model's learning efficiency and generalization.

As shown in Figure 3(a), the training accuracy remains consistently high across all experimental ratios, ranging from 93.58% to 95.83%. This indicates that the Naive Bayes Classifier (NBC), when combined with the TF–IDF feature extraction method,

effectively learns the emotional features within the dataset. In contrast, the testing accuracy shows smaller fluctuations between 67.17% and 68.86%, suggesting that while the model generalizes well to unseen data, its accuracy on the test set is slightly lower than on the training set. The highest testing accuracy (68.86%) is achieved at the 80:20 ratio, which represents the optimal balance between learning capacity and evaluation stability. This ratio provides the model with sufficient data to capture meaningful emotion-related patterns, while maintaining enough unseen samples to test its generalization. Beyond this point, such as in the 90:10 ratio, the accuracy improvement becomes marginal (68.59%), since reducing the test size limits the robustness of the evaluation. Overall, Figure 3(a) indicates balanced model performance, with a small and stable gap between training and test accuracy, suggesting no severe overfitting. The model maintains consistent generalization across different proportions of data, indicating good stability in emotion detection for Indonesian-language tweets.

Figure 3(b) shows the variation in training loss and testing loss (calculated as 1 – accuracy) across the same experimental ratios. The training loss remains consistently low, ranging from 4.17% to 6.42%, indicating that the model successfully minimizes prediction errors during training. The slight downward trend as the training data increases suggests that additional samples help the classifier refine its probabilistic boundaries, thereby reducing misclassification. Meanwhile, the testing loss fluctuates between 31.14% and 32.83%, showing minor variations across different ratios. These stable loss values indicate that the model maintains reliable performance across different training-to-testing ratios. The consistency between loss and accuracy patterns also supports the observation that the model avoids significant overfitting while retaining generalization across unseen data. Taken together, the results from Figures 3(a) and 3(b) demonstrate that the TF–IDF and Naive Bayes combination achieves high training performance and stable testing results across varying dataset ratios. The minimal deviation between accuracy and loss metrics indicates that the model is both efficient and robust, making it suitable for emotion classification in Indonesian-language social media data, where linguistic variability and informal expressions are common. Overall, the results demonstrate that the combination of TF–IDF feature extraction and a Naive Bayes Classifier produces stable, consistent classification accuracy across different dataset ratios. The small gap between the four scenarios suggests that the model is relatively robust and can handle variations in data size, making it suitable for emotion detection in Indonesian-language social media data.

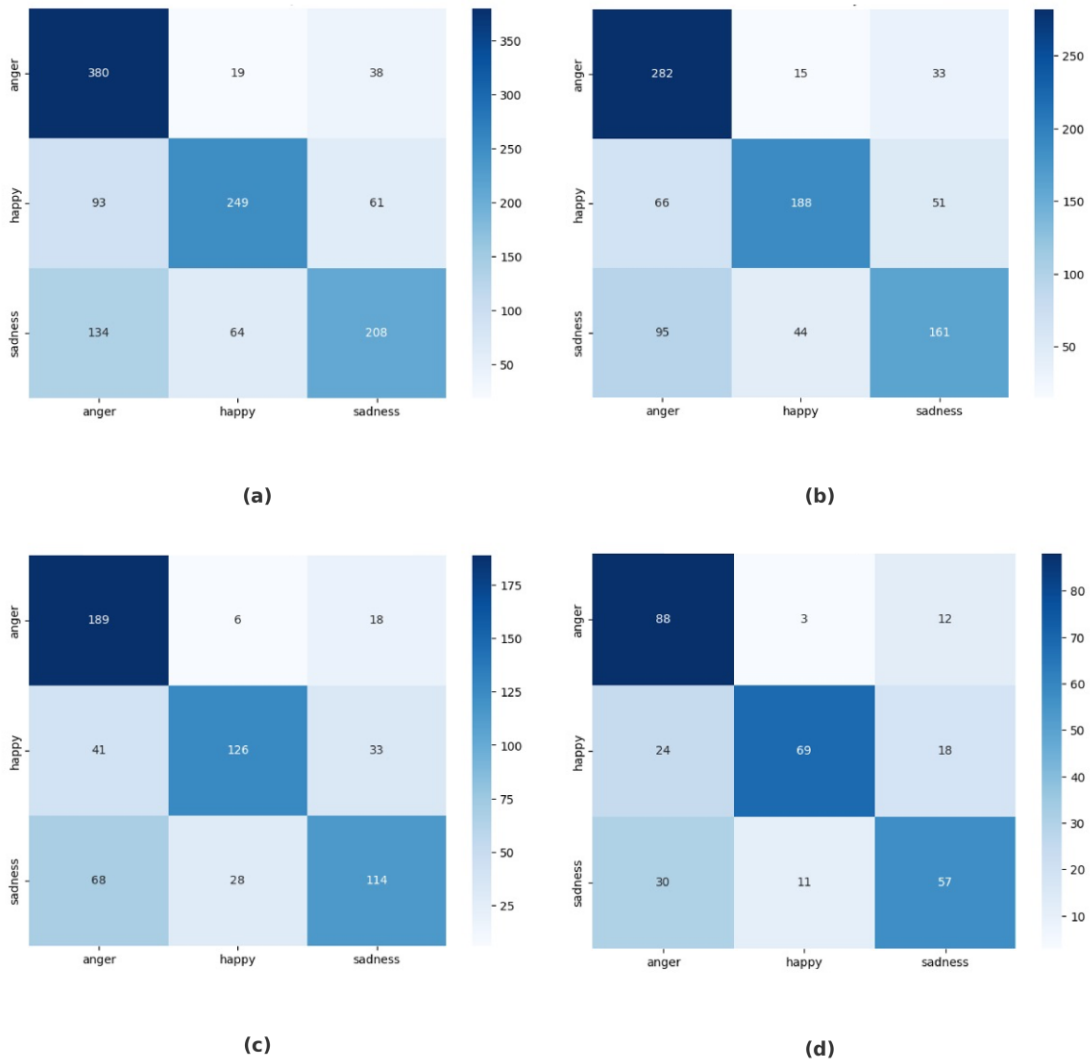


Fig.4. Confusion Matrices For Emotion Classification At Different Training-Testing Ratios (a)60:40, (b)70:40, (c)80:20, (d)90:10.

DOI: http://dx.doi.org/10.25139/ijair.v7i1.11156

The distribution of correctly and incorrectly classified samples across three emotion categories (anger, happiness, and sadness) for four experimental data ratios. The confusion matrices provide a visual representation of how well the Naive Bayes Classifier with TF–IDF feature extraction distinguishes between emotional classes under different proportions of training and testing data, shown as Fig.4. In Figure 4(a), corresponding to the 60:40 ratio, the model demonstrates a relatively high ability to identify the *anger* class, with 380 samples correctly classified. However, misclassifications still occur, particularly between *sadness* and *happiness, which often overlap* in informal Indonesian text. The happy class achieves 249 correct predictions, while the sadness class records 208, indicating moderate precision across the two classes. When the training data proportion increases to 70% as shown in Figure 4(b), the model exhibits improved overall performance. The *anger* class remains the most accurately identified with 282 correct predictions, while *happy* reaches 188 and *sadness* 161.

This improvement suggests that adding more training samples helps the model learn emotional patterns more effectively, aligning with observations from previous studies emphasizing the benefit of richer training corpora in emotion classification tasks [8]. Figure 4(c) shows the 80:20 split, in which the model achieves its optimal balance between accuracy and generalization. In this configuration, the classifier performs most consistently across all classes, correctly predicting 189 anger samples, 126 happy samples, and 114 sadness samples. The reduction in misclassification rates confirms that the model generalizes well without overfitting the training data, consistent with earlier findings based on accuracy and loss trends. At the 90:10 ratio, shown in Figure 4(d), the model still maintains a stable classification pattern, but the number of correctly identified samples decreases slightly. The anger class remains dominant with 88 correct predictions, whereas happiness and sadness achieve 69 and 57, respectively. Although training accuracy continues to rise slightly, the limited test data size likely limits the model's ability to fully validate its generalisation. Across all ratios, the anger class yields the highest number of correct predictions.

This behaviour indicates that emotional cues associated with anger—such as the presence of strong negative words or exclamatory expressions—are more distinct and thus easier for the classifier to detect. In contrast, happiness and sadness exhibit more ambiguity due to overlapping contextual and linguistic features in Indonesian tweets, a phenomenon also discussed in emotion lexicon research [6]. Overall, the confusion matrices in Fig. 4 demonstrate that the Naive Bayes Classifier with TF–IDF effectively captures emotional distinctions in textual data. The 80:20 ratio emerges as the most balanced configuration, offering high classification accuracy with minimal overfitting. These results reaffirm the robustness of Naive Bayes for multilingual emotion detection, particularly in resource-limited languages such as Indonesian.

TABLE I
TESTING RESULT

Testing scenario	Accuracy	Recall			precession		
		anger	happy	sad	anger	happy	sad
60:40	67,17%	0,87	0,62	0,51	0,63	0,75	0,68
70:30	67,48%	0,85	0,62	0,54	0,64	0,76	0,66
80:20	68,86%	0,89	0,63	0,54	0,63	0,79	0,69
90:10	68,58%	0,85	0,62	0,58	0,62	0,83	0,66

TABLE II
RECALL, PRECISION AND F1-SCORE RESULT

Scenario	Accuracy	Recall (anger, happy, sad)	Precision (anger, happy, sad)	F1-score (anger, happy, sad)
60.40.00	67.17%	0.87 / 0.62 / 0.51	0.63 / 0.75 / 0.68	0.73 / 0.68 / 0.58
70.30.00	67.48%	0.85 / 0.62 / 0.54	0.64 / 0.76 / 0.66	0.73 / 0.68 / 0.59
80.20.00	68.86%	0.89 / 0.63 / 0.54	0.63 / 0.79 / 0.69	0.74 / 0.70 / 0.61
90.10.00	68.58%	0.85 / 0.62 / 0.58	0.62 / 0.83 / 0.66	0.72 / 0.71 / 0.61

The experimental evaluation was conducted using four testing scenarios with different data split ratios: 60:40, 70:30, 80:20, and 90:10. Each scenario was assessed based on three primary performance metrics—Accuracy, Recall, and Precision—to determine the classification performance of the Naive Bayes Classifier combined with TF–IDF feature extraction. The detailed results are summarized in Table 1. Model results in its highest accuracy of 68.86% under the 80:20 ratio (Scenario 3), indicating that this configuration provides the best balance between learning capacity and testing stability. The Recall values for the *anger* and *happy* emotion classes also reached their peak in this scenario, with 0.89 and 0.63, respectively. Meanwhile, the *sadness* class achieved its highest recall (0.58) at the 90:10 ratio (Scenario 4).

This suggests that a slightly larger training dataset can help the model capture features of emotions that are less frequently expressed, such as sadness. In terms of precision, the highest values vary across emotion classes. The angry class achieved the highest precision (0.64) at the 70:30 ratio (Scenario 2), while the happy class achieved the highest precision (0.83) at the 90:10 ratio (Scenario 4). For the sadness class, the optimal precision (0.69) was observed at the 80:20 ratio (Scenario 3). These variations suggest that the model's ability to predict certain emotions correctly depends on the distribution of the training data and the emotional distinctiveness within each subset. Overall, Scenario 3 (80:20) demonstrates the most consistent and balanced performance across all metrics. It achieved high accuracy and recall while maintaining stable precision across classes, confirming its suitability as the optimal configuration for this model.

DOI: http://dx.doi.org/10.25139/ijair.v7i1.11156

This ratio allows the classifier to generalize effectively without overfitting, reflecting robust learning of emotional features in Indonesian-language text. Similar findings have been reported in previous studies on text-based emotion detection, where moderate data ratios often yield the best generalization performance. In contrast, Scenario 1 (60:40) yielded the lowest accuracy (67.17%), suggesting reduced generalization due to limited training data. This scenario also resulted in the lowest Recall and Precision for certain classes, such as anger (Recall = 0.87) and happy (Precision = 0.75). The overall performance trend indicates that increasing the proportion of training data improves the classifier's ability to identify emotional patterns and reduces misclassification errors. From these results, it can be concluded that the Naive Bayes Classifier combined with TF–IDF effectively classifies emotions in Indonesian Twitter comments, with an 80:20 split providing the best trade-off between accuracy, recall, and precision. The model exhibits stable, interpretable behaviour across different configurations, validating its potential for practical emotion-detection applications in social media analytics.

In addition to Accuracy, Recall, and Precision, Table 2 presents the F1-score for each emotion class across all testing scenarios to provide a more balanced evaluation of model performance. The F1-score results further reinforce the conclusion that the 80:20 ratio offers the most stable configuration. In this scenario, the anger class achieves an F1-score of **0.74**, the happy class achieves **0.70**, and the sadness class achieves **0.61**, representing the most consistent performance across all emotion categories. Conversely, the 60:40 and 70:30 ratios show noticeably lower F1-scores for the sadness class, indicating reduced ability to correctly detect this less dominant emotion when the training set is smaller. Meanwhile, the 90:10 scenario produces higher Precision for some classes but yields F1-scores comparable to those of the 80:20 split due to reduced recall from the smaller testing set. This pattern suggests that while extreme training ratios may boost Precision, they do not always translate into better-balanced performance. Overall, the F1-score analysis in Table 2 confirms that the 80:20 split provides the best trade-off between recall and precision across all classes. These results emphasize that a moderate amount of training data enables the classifier to learn emotional patterns effectively while still maintaining robust performance on unseen data.

IV. CONCLUSION

This study demonstrated that combining the Naive Bayes Classifier with TF–IDF feature extraction effectively detects emotions in Indonesian Twitter comments across the three selected categories—anger, happiness, and sadness. Among the four testing ratios applied during model development, the 80:20 configuration yielded the best overall performance, achieving 68.86% accuracy with balanced recall and precision. These results indicate that, despite its simplicity, Naive Bayes remains a dependable and computationally efficient method for emotion classification in textual data. An important aspect to emphasize is the difference in performance between the public and validation datasets.

The model in this study was trained and tested on a publicly available English-language emotion dataset, split into training and test sets. Meanwhile, the validation dataset consisted of Indonesian tweets collected through scraping. Because no publicly available Indonesian emotion dataset exists for the three target classes, the validation stage necessarily relies on scraped data that contains higher linguistic variability, informal expressions, and potential label noise. For this reason, obtaining validation accuracy below 80% is entirely expected and primarily reflects differences in dataset characteristics rather than flaws in the model or its implementation. The discrepancy is therefore a natural outcome of comparing a well-structured public dataset with raw, real-world Indonesian social media data.

Looking ahead, several improvements can be explored before transitioning to deep learning approaches. Accuracy can be enhanced by optimizing the hyperparameters of Naive Bayes and TF–IDF, refining the handling of class imbalance with methods such as oversampling or SMOTE, and incorporating weak supervision techniques, for instance by using tools like VADER as a preliminary labelling mechanism to enrich the training dataset. Additional refinements in text pre-processing—such as more advanced handling of emojis, slang normalization, and context-aware cleaning—may also reduce noise and improve the model's ability to generalize to Indonesian validation data. Once these conventional improvements have been sufficiently explored, adopting deep learning models such as LSTM, BERT, or IndoBERT may yield further gains by capturing deeper contextual information. Finally, applying k-fold cross-validation in future studies would yield more stable and unbiased performance estimates compared to a single train–test split.

REFERENCES

[1] E. Ford, S. Shepherd, K. Jones, and L. Hassan, "Toward an ethical framework for the text mining of social media for health research: a systematic review," *Frontiers in Digital Health*, vol. 2, pp. 1–19, 2021, doi: 10.3389/fdgth.2020.592237.
[2] DataReportal, “Digital 2023: Indonesia.” 2023. [Online]. Available: https://datareportal.com/reports/digital-2023-indonesia
[3] K. Arifin and S. I. Al-Idrus, “Klasifikasi emosi pengguna twitter terhadap bakal calon presiden pada pemilu 2024 menggunakan algoritma naïve bayes,” *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)*, vol. 23, no. 1, p. 37, 2024, doi: 10.53513/jis.v23i1.9558.
[4] A. Pratama and L. Findawati, "Analysis of hate speech on Platform X in Indonesian online discourse," *Journal of Social Media and Communication Studies*, vol. 5, no. 1, pp. 45–54, 2024.
[5] P. Ekman, "Basic emotions," in *Handbook of cognition and emotion*, T. Dalgleish and M. Power, Eds., John Wiley & Sons, 1999.
[6] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–19, 2021, doi: 10.1007/s13278-021-00776-6.

[7] A. S. Aribowo and S. Khomsah, "Implementation of text mining for emotion detection using the lexicon method (case study: Tweets about covid-19)," *Telematika*, vol. 18, no. 1, p. 49, 2021, doi: 10.31315/telematika.v18i1.4341.
[8] K. Machová, M. Szabóova, J. Paralič, and J. Mičko, "Detection of emotion by text analysis using machine learning," *Frontiers in Psychology*, vol. 14, p. 1190326, 2023, doi: 10.3389/fpsyg.2023.1190326.
[9] M. Z. Anbari and B. Sugiantoro, "Studi komparasi metode analisis sentimen naïve bayes, SVM, dan logistic regression pada piala dunia 2022," *Mikrotik: Jurnal Ilmiah Informatika*, vol. 7, no. April, pp. 688–695, 2023, doi: 10.30865/mib.v7i2.5383.
[10] R. Krishnan and M. Elayidom, "Emotion detection using Naive Bayes classifier," *Procedia Computer Science*, vol. 115, pp. 372–379, 2017, doi: 10.1016/j.procs.2017.09.123.
[11] A. Khalil, "Comparison of SVM and naive bayes algorithms for emotion detection," *International Journal of Artificial Intelligence Research*, vol. 3, no. 2, pp. 78–85, 2022.
[12] S. Azmin, "Emotion detection from bangla text corpus using naive bayes classifier," in *2019 4th international conference on electrical information and communication technology (EICT)*, 2019, pp. 1–5. doi: 10.1109/EICT48899.2019.9068797.
[13] R. Sankar, P. Sharma, and V. Kumar, "Sentiment analysis on Twitter data for depression detection," *Journal of Machine Learning Applications*, vol. 9, no. 1, pp. 33–42, 2024.
[14] M. Anbari and B. Sugiantoro, "Improving text classification performance using TF-IDF weighting method," *Journal of Information Systems and Technology*, vol. 11, no. 2, pp. 65–72, 2023.
[15] P. Widyantara, A. Wibawa, and C. Pramartha, "Analisis sentimen pada teks berbahasa bali menggunakan metode multinomial naive bayes dengan TF-IDF dan BoW," *Journal of Language Technology and Computing*, vol. 2, no. November, pp. 37–46, 2023.
[16] R. Mazya, Suryani, and R. Pratama, "Comparative study of feature extraction using TF-IDF and Bag of Words," *International Journal of Data Science and Technology*, vol. 8, no. 4, pp. 121–128, 2022.
[17] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis sentimen pengguna twitter terhadap polemik persepakbolaan indonesia menggunakan pembobotan TF-IDF dan k-nearest neighbor," *Journal of Intelligent System and Computation*, vol. 1, no. 1, pp. 43–49, 2019, doi: 10.52985/insyst.v1i1.36.
[18] S. Mazya, P. Tyas, B. S. Rintyarna, and W. Suharso, "The impact of feature extraction to naïve bayes based sentiment analysis on review dataset of indihome services," *Journal of Informatics and Computational Science*, pp. 1–10, 2022.
[19] T. Ridwansyah, "Implementasi text mining terhadap analisis sentimen masyarakat dunia di twitter terhadap kota medan menggunakan k-fold cross validation dan naïve bayes classifier," *Jurnal Teknologi dan Sains Komputer*, vol. 2, no. 5, pp. 178–185, 2022.
[20] A. Sudrajat, "Penerapan metode naïve bayes untuk menentukan penilaian kinerja karyawan PT. Sinergi guna solusindo," *Jurnal Teknologi dan Sains Komputer*, vol. 99, no. 99, pp. 1596–1606, 2022.
[21] C. H. Pratama and Y. Findawati, "Klasifikasi hate speech dan emosi dalam teks berbahasa indonesia pada pengguna twitter menggunakan metode naïve bayes classifier," *Indonesian Journal of Applied Technology*, vol. 1, no. 3, p. 10, 2024, doi: 10.47134/ijat.v1i3.3105.
[22] M. Muntean and F.-D. Militaru, "Metrics for evaluating classification algorithms," in *Education, research and business technologies*, C. Ciurea, P. Pocatilu, and F. G. Filip, Eds., Springer Nature Singapore, 2023, pp. 307–317.
[23] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Artificial intelligence application in networks and systems*, R. Silhavy and P. Silhavy, Eds., Springer International Publishing, 2023, pp. 15–25.
[24] A. S. Aribowo and N. Khomsah, "Emotion analysis in Indonesian social media comments using EmoLex lexicon," *Indonesian Journal of Computational Linguistics*, vol. 6, no. 2, pp. 112–120, 2021.
[25] F. Khalil Aljwari, "Emotion detection in arabic text using machine learning methods," *International Journal of Information System and Computer Science (IJISCS)*, vol. 6, no. 5, pp. 175–185, 2022.