MJI
Media Jurnal Informatika
Vol. 17, No. 2, December 2025
p-issn : 2088-2114, e-issn : 2477-2542

Emotion Detection in Indonesian Text Using the Logistic Regression Method

Erfian Junianto[a,1,*], Mila Puspitasari[b,2], Salman Ilyas Zakaria [b,3], Toni Arifin[a,4], Ignatius Wiseto Prasetyo Agung[a,5]

[a] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya, Bandung, Indonesia 40282
[b] Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya, Bandung, Indonesia 40282
[1] erfian.ejn@ars.ac.id*; [2] mileuups14@gmail.com; [3] salmanzakaria38@gmail.com; [4] toni.arifin@ars.ac.id; [5] wiseto.agung@ars.ac.id;
* corresponding author

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Deteksi Emosi
Logistic regression
Ensemble Bagging
Text Mining
Data Tekstual

ABSTRACT

Emotion detection in Indonesian text has become a crucial topic in the advancement of human–computer interaction and sentiment analysis on digital platforms. Despite its importance, challenges arise from the linguistic complexity and frequent use of slang in Indonesian text. This study aims to evaluate the performance of three classification models—Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes—in detecting emotions from Indonesian text. The dataset comprises 1,000 texts categorized into four emotions: happy, sad, angry, and fear. Preprocessing steps included slang normalization, text cleaning, tokenization, stopword removal, and stemming, followed by TF-IDF weighting. Each model was trained and further optimized using ensemble bagging to improve classification performance. The optimized Logistic Regression model achieved the best performance, with an accuracy of 89%, precision of 0.90, recall of 0.89, F1-score of 0.89, and an average ROC-AUC score of 0.98. Both KNN and Naive Bayes models reached 81% accuracy after optimization, but their overall performance remained lower than Logistic Regression. These results indicate that Logistic Regression provides the most consistent and reliable performance for emotion detection in Indonesian text, while the ensemble approach mainly contributes to improving prediction stability and yields more substantial benefits for weaker classifiers. This study contributes to the development of emotion analysis models for Indonesian text, supporting applications in social computing and affective computing.

1. Introduction

Research on human emotions has long been a central focus across various disciplines, such as cognitive science, psychology, and, more recently, computer science with the rapid growth of social media. Understanding emotions is essential for advancing human–computer interaction and for exploring social trends in diverse fields, particularly those related to psychological issues. Emotions play a significant role in daily human life, influencing social relationships, memory, and even decision-making processes [1].

Text is one of the primary media used for communication and information delivery. Beyond conveying information, text can also express emotions [2]. Emotional textual data are increasingly abundant with the growing use of social media. This trend necessitates the development of more efficient methods for detecting emotions in Indonesian text [3]. Moreover, textual data often exhibit diverse features, such as variations in writing style, the use of slang, and dialectal differences, all of which can affect the accuracy of emotion detection. Therefore, this study also emphasizes the importance of selecting appropriate feature engineering techniques to address such diversity and to enhance the performance of emotion detection models [4].

Text mining is a technology used to discover useful knowledge within collections of text documents, enabling the identification of trends, patterns, or similarities in natural language texts that serve specific purposes [5]. Text mining is also a process of extracting valid and applicable knowledge from various

documents and utilizing this knowledge to better organize information for future reference [6]. Text within documents consists of various types of words, such as prepositions, conjunctions, pronouns, adjectives, and others. Some of these words cannot be used as document indices because their occurrences are not specific or unique to particular documents [7]. Through text mining, it becomes possible to extract and uncover valuable information from textual data [8].

Emotion classification methods are used to detect emotions, where emotion classes are determined based on the analyzed text [9]. Emotion detection in Indonesian text is one of the challenges in text mining that requires appropriate approaches to achieve accurate results. Emotions expressed in text can be utilized to understand public sentiment, analyze social media, and support other applications related to human interaction [3]. In the process of emotion identification, several physiological characteristics can also be employed, such as voice, facial expressions, hand gestures, body movements, heartbeat, blood pressure, as well as information obtained from textual data.

Emotion detection in Indonesian text is increasingly important with the growing use of the language across digital platforms such as social media, e-commerce, and online forums. The choice of Indonesian text in this study is based on the significant growth of internet and social media users in Indonesia, which provides great opportunities for text-based emotion analysis in a local context [10].

Indonesia's rapid growth in internet usage has led to a substantial increase in user-generated textual content on social media, online forums, and digital communication platforms. These texts are characterized by informal language, extensive slang usage, abbreviations, code-mixing, and relatively simple grammatical structures. Such linguistic properties present unique challenges for emotion detection, particularly in low-resource language settings where standardized lexical resources are limited. Consequently, robust and interpretable machine learning models that can effectively handle sparse features and lexical variation—such as Logistic Regression—are highly relevant for Indonesian emotion classification tasks [10].

Table 1. Internet Penetration Rate

Year	Penetration
2018	64,80%
2020	73,70%
2022	77,01%
2023	78,19%
2024	79,50%

Table 1 presents the development of internet penetration rates in Indonesia from 2018 to 2024. The data illustrate a consistent upward trend, with internet penetration reaching 79.50% in 2024. The Indonesian language offers several advantages in terms of its relatively simpler structure compared to English. The absence of verb conjugation based on tenses or subjects facilitates text processing and the application of modeling techniques [11]. This structural simplicity can also reduce the complexity of detecting emotions expressed explicitly in text.

However, emotion detection in Indonesian text also faces unique challenges. First, emotional expressions in Indonesian are often more formal and less explicit compared to English, making emotion classification more difficult [12]. Additionally, the limited vocabulary of Indonesian in representing diverse emotional nuances often relies on loanwords or local dialects, which are not always consistent. Another challenge is the scarcity of adequate Indonesian-language datasets for training emotion detection models, which often necessitates manual annotation [3].

Previous studies have evaluated the performance of Deep Learning methods in detecting emotions in social media text using several datasets, including Semeval, WASSA, Tweet Pemilu, and Crowdflower. The experimental results demonstrated that Deep Learning is among the most effective methods for emotion classification. On the Semeval dataset, the CNN architecture achieved the highest accuracy of 81.64%. Meanwhile, for the WASSA dataset, CNN, MLP, and GRU methods showed comparable performance. On the Tweet Pemilu dataset, LSTM and GRU achieved the highest accuracy. For the Crowdflower dataset, the LSTM, RNN, and GRU methods yielded the best performance, with the highest accuracies of 92.33% for LSTM and 92.30% for both RNN and GRU [13].

Subsequent research compared the performance of Logistic Regression and K-Nearest Neighbors (KNN) in text classification using TF-IDF and hyperparameter tuning. TF-IDF significantly improved the

performance of KNN, with accuracy and F1-score increases of up to 48.4% and 54.84%, respectively, whereas for Logistic Regression, only precision improved by 2.985%. The combination of TF-IDF and hyperparameter tuning yielded the best results for Logistic Regression, with an accuracy of 65% and an F1-score of 66% [14].

Another study focused on classification methods and feature extraction for sentiment analysis, with most datasets derived from Twitter. Commonly used algorithms included Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, and lexicon-based approaches. The results showed that Logistic Regression achieved the highest accuracy at 93.60%, followed by the lexicon-based approach with 92%, while Naïve Bayes, SVM, Random Forest, and K-Means achieved 88.20%, 85.50%, 81%, and 84.16%, respectively. The superior performance of Logistic Regression and lexicon-based methods was likely influenced by optimal feature extraction and effective dataset management [15].

Previous studies also compared the performance of machine learning and deep learning models in sentiment analysis of Shopee customer reviews using a dataset of 6,002 comments. The machine learning models tested included Logistic Regression, Naïve Bayes, and Multinomial Naïve Bayes, with Logistic Regression achieving the highest accuracy at 94.58%. In contrast, for deep learning models, BERT achieved an accuracy of 92.83%, while Multilingual BERT produced the best results with 97.41% accuracy. These findings suggest that deep learning models demonstrate superior accuracy compared to machine learning models in customer sentiment analysis [16]–[19].

Other studies have also revealed that Logistic Regression achieved the highest accuracy in detecting emotions in English text [20], particularly when combined with ensemble bagging techniques [21]. The Indonesian language, however, presents unique characteristics compared to English, such as a more limited vocabulary for describing emotions and the tendency to use more formal expressions. These factors pose challenges for emotion detection in Indonesian text. Therefore, Logistic Regression is considered more suitable for text classification tasks in both Indonesian and English, due to its ability to handle high-dimensional and sparse data, which are common in text representations using TF-IDF. Models such as KNN, although effective for non-linear data distributions, are generally less efficient for textual data as they require more memory and computational time, especially when applied to large-scale datasets [22], [23].

Accordingly, this study aims to analyze the effectiveness of Logistic Regression in detecting emotions in Indonesian text. The main focus is on applying ensemble bagging techniques and feature engineering to improve the accuracy of emotion detection models, taking into account the simpler grammatical structure of Indonesian compared to English. In addition, this research incorporates an extra preprocessing stage for handling slang words, which are commonly found in Indonesian text. The study also seeks to identify key challenges, such as the tendency of emotional expressions in Indonesian to be more formal and less explicit, as well as the limitations in vocabulary that may influence detection outcomes.

2. Method

Figure 1 illustrates the stages or processes of the proposed model for emotion detection using the Ensemble Bagging approach. The diagram outlines the sequential steps performed, starting from the input of textual data to the final output of detected emotions.

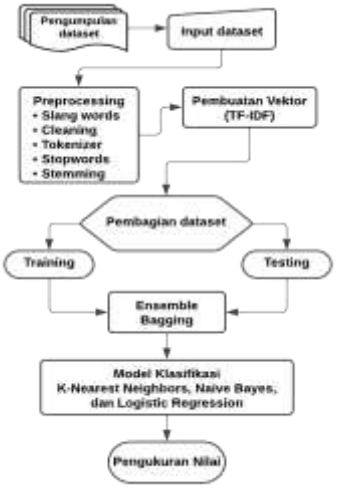

Fig. 1. Research Method

2.1 Dataset Collection

In this study, the data were obtained from a GitHub repository providing a dataset to support emotion detection analysis in Indonesian text. The dataset consists of four emotion categories: happy, sad, angry, and fear. Each emotion category contains 250 text samples, with the total dataset comprising 1,000 samples. The dataset size is presented in Table 2.

Table 2. Dataset Size

Class	Total
Happy	250
Sad	250
Angry	250
Fear	250

Table 2 shows the distribution of the four emotions in the dataset: happy, sad, angry, and fear [24]. Each emotion category contains 250 samples, resulting in a total of 1,000 samples. This balanced distribution is essential to prevent bias toward any particular emotion, ensuring that the analysis and the developed model can provide more accurate and balanced results in identifying or classifying emotions.

2.2 Preprocessing

Text preprocessing is the process of preparing text for analysis by converting unstructured data into structured data. Typically, structured data are represented in numerical form [25]. The preprocessing stages in this study include handling slang words, cleaning and lowercasing, tokenization, stopword removal, and stemming. Slang words refer to language commonly used in informal conversations by teenagers or young adults, whether in the United States, the United Kingdom, or Indonesia. The slang words stage is performed prior to cleaning, where non-standard or slang words are converted into standard words. For example, "knp" is transformed into "kenapa," and "jg" becomes "juga," to ensure that the text is more formal and conforms to standard grammar [26]. This process is illustrated in Table 3.

Table 3. Slang Words

Text	Output
Gk ada yang mw ngerjain tp sy harus lakuin karena udah telat banget, dan harus segera selesai.	Tidak ada yang mau mengerjakan tapi saya harus lakukan karena sudah telat banget, dan harus segera selesai.

The cleaning stage aims to reduce noise in the data by removing account names, numbers, "RT" tags, hashtags, duplicates, emoticons, punctuation, and hyperlinks. The lowercasing stage converts all text in the documents into a standard format, typically lowercase letters from 'a' to 'z' [27]. This process is shown in Table 4.

Table 4. Slang Words

Text	Output
Tidak ada yang mau mengerjakan tapi saya harus lakukan karena sudah telat banget, dan harus segera selesai.	tidak ada yang mau mengerjakan tapi saya harus lakukan karena sudah telat banget, dan harus segera selesai

The next stage, tokenization, separates sentences into individual words, which are then arranged in an array format [27], as shown in Table 5.

Table 5. Tokenizer

Text	Output
tidak ada yang mau mengerjakan tapi saya harus lakukan karena sudah telat banget, dan harus segera selesai	["tidak", "ada", "yang", "mau", "mengerjakan", "tapi", "saya", "harus", "lakukan", "karena", "sudah", "telat", "banget", "dan", "harus", "segera", "selesai"]

The subsequent stage is stopword removal, which eliminates common words that frequently appear but do not affect sentiment [27], as presented in Table 6.

Table 6. Stopwords

Text	Output
["tidak", "ada", "yang", "mau", "mengerjakan", "tapi", "saya", "harus", "lakukan", "karena", "sudah", "telat", "banget", "dan", "harus", "segera", "selesai"]	[“tidak”, “mau”, “mengerjakan”, “harus”, “lakukan”, “telat”, “banget”, “harus”, “segera”, “selesai”]

The final stage is stemming, the process of converting words to their root form by removing prefixes, infixes, or suffixes. The Sastrawi library was used for stemming in this study [27], as illustrated in Table 7.

Table 7. Stemming

Text	Output
tidak mau mengerjakan harus lakukan telat banget harus segera selesai	tidak mau kerja harus laku telat banget harus segera selesai

2.3 Vector Creation

After preprocessing, the next step is to convert the collection of words into vectors using TF-IDF weighting [8]. Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical method used in natural language processing and information retrieval systems to assess the importance of a term (keyword) within a document relative to a larger collection of documents. TF-IDF generates a score that reflects how significant a term is in a specific document compared to other documents. This score can be used to measure and compare the relevance of documents within a retrieval system. Documents with higher TF-IDF scores for a given term are generally considered more relevant to user queries containing that term. TF-IDF is a widely used method for enhancing accuracy and relevance in information retrieval [28].

The dataset was divided using a 90:10 ratio. From the total dataset of 1,000 samples, this split results in 900 samples for training and 100 samples for testing. The purpose of this division is to ensure that the model can be evaluated on previously unseen data, allowing its performance to be assessed objectively [29].

2.4 Classification Model

After the dataset splitting process, the next step is to apply the classification model using a combination of machine learning algorithms, namely K-Nearest Neighbors (K-NN), Logistic Regression, and Naive Bayes, through the ensemble bagging approach. In this study, the results of the three algorithms are integrated at the testing stage. Each algorithm first processes the test data separately. The predictions generated by each algorithm are then combined using the ensemble bagging technique, producing a final prediction that is more stable and accurate.

$$\widehat{f_{bag}} = \widehat{f_1}(X) + \widehat{f_2}(X) + \cdots + \widehat{f_b}(X) \quad (1)$$

Equation 1 illustrates the ensemble process in the bagging method, where the final prediction $\widehat{f_{bag}}$ is obtained by summing or aggregating the predictions $\widehat{f_1}(X) + \widehat{f_2}(X) + \cdots + \widehat{f_b}(X)$ generated by several base learners. The measurement process evaluates the ability of the learning algorithm models to solve problems or manage data effectively [19].

2.5 Measurement Value

This evaluation utilizes true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values arranged in a confusion matrix to determine the performance of the machine learning algorithms. The data from the confusion matrix are used to calculate accuracy, precision, recall, and F1-score, which serve as indicators of the applied algorithm's performance [30]. The accuracy score reflects the proportion of correct predictions relative to the total number of predictions, with a maximum value of 1 and a minimum value of 0. The formulas for calculating accuracy, precision, recall, and F1-score are presented in Equations 2, 3, 4, and 5 [31].

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1\ score = \frac{2(Precision \cdot Recall)}{Precision+Recall} \quad (5)$$

3. Results and Discussion

This section presents the results and discussion of the conducted research. The main aspects discussed include Research Data, Data Preprocessing, Weighting Results, Dataset Splitting, Classification Modeling, and Evaluation Metrics.

3.1. Research Data

The research data were obtained through a web scraping process from a GitHub repository containing the relevant dataset, including a slang dictionary used in this study, as shown in Table 8.

Table 8. Sample Texts from the Dataset in the GitHub Repository

No	Text
1	Ga harus ngomel-ngomel,bisa?!
2	lah ko ketawa masalah ?
3	Makan Sehat Hidup Senang
4	pas gua udah mulai suka , udah membuka hati lagi lu malah ngecewain
5	ku sedih banget bulan itu aku udah uts jadi gak bisa ke jakarta :((huhuhu pengen banget ktmu kalex
6	Aten rindu saya sampai letak gmbar berdua dengan saya di wechat
7	GUE GA PERDULI LO MAU NGOMONG APAAN
8	Pinter boong, suka ngeles, pinter cari muka, otak kecil , yah itu si "budakkecikbalita"
9	smtime bkn krn kebohongan utk membenci ssorg, tp krn sedih menerima kenyataan bahwa ia tak bisa lg dipercaya:')
10	Jangan sedih bila sekarang masih dipandang sebelah mata, buktikan bahwa anda layak mendapatkan kedua matanya.

Table 8 illustrates sample texts from the dataset employed in this study, consisting of 1,000 Indonesian-language texts that had been annotated with emotion labels. Each text is categorized into several types of emotions, such as anger, sadness, and fear. For instance, under the anger label, an example text is: “SUSAH NGOMONG SAMA ORANG YANG GA TAU DIRI.” Meanwhile, under the sadness label, a sample text is: “macam mw balas tweet klau sy x pandai english;. haiszz #sedih.” Each emotion category covers texts with diverse contexts, ranging from expressions of frustration and loss to fear regarding particular situations. The dataset was sourced from the GitHub repository that compiles Indonesian-language texts related to emotional expression.

Since the dataset contained informal and non-standardized language, preprocessing was required to enhance data quality. One critical step in this process was slang word replacement. This was accomplished by using the slang dictionary included within the dataset to substitute non-standard words or abbreviations with

their standardized forms, such as converting “sy” to “saya” and “gk” to “tidak.” This step was essential to ensure textual consistency and to facilitate subsequent model processing. The preprocessing prepared the dataset to be optimally utilized for emotion detection in Indonesian texts, as presented in Table 9.

Table 9. Sample of the Slang Dictionary

Slang	Standard Word
ad	ada
aj	aja
atw	atau
bgt	banget
bnyk	banyak
gk	gak
iy	iya
jd	jadi
kyk	kayak
lg	lagi

Table 9 displays 10 word pairs out of a total of 542 pairs as a sample from the dataset. For example, slang terms such as “ad” are transformed into “ada”, “btw” into “banget”, and “bnyk” into “banyak.” This slang dictionary plays a crucial role in the preprocessing stage to ensure that the texts follow a more standardized format, making them easier to process by computational algorithms. The dictionary was employed for text normalization, i.e., converting informal or non-standard words into their formal equivalents. Such normalization is vital for maintaining data consistency in text processing, which in turn facilitates analysis and improves model accuracy, particularly in tasks such as emotion detection. This process significantly contributes to enhancing the performance of emotion detection models and other text analysis applications by reducing ambiguity and linguistic variation.

3.2. *Data Preprocessing*

The comments presented in Table 10, which are intended for classification or prediction purposes, were first processed through a text preprocessing pipeline. This stage included handling slang or non-standard words, removing stopwords, eliminating punctuation marks, converting all text to lowercase, and applying stemming to each word [32]. At this stage, Indonesian-language comments were processed through several sequential steps. First, informal or slang words were cleaned and converted into their standardized forms. Next, all text was transformed into lowercase. The text was then tokenized into individual words. The process continued with the removal of common words that do not carry significant meaning, such as “dan” (and), “atau” (or), “ke” (to), and “di” (in) (stopword removal). Finally, stemming was applied to reduce words to their root forms [33].

Table 10. Preprocessing

No	Text	Preprocessing
1	knp ka?? aku jg gk denger siaran trakhir kk lg. #sedih. Mksih ya ka udh jd tmn aku di stiap kk lg siaran	Text
2	kenapa ka?? aku juga tidak denger siaran terakhir kak lagi . # sedih . makasih ya ka sudah jadi teman aku di setiap kak lagi siaran	Slang Words
3	kenapa ka aku juga tidak denger siaran terakhir kak lagi sedih makasih ya ka sudah jadi teman aku di setiap kak lagi siaran	Cleaning & Lowercase
4	[kenapa, ka, aku, juga, tidak, denger, siaran, terakhir, kak, lagi, sedih, makasih, ya, ka, sudah, jadi, teman, aku, di, setiap, kak, lagi, siaran]	Tokenizer
5	[ka, denger, siaran, terakhir, kak, sedih, makasih, ka, jadi, teman, kak, siaran]	Stopword Removal
6	[ka, denger, siar, akhir, kak, sedih, makasih, ka, jadi, teman, kak, siar]	Stemming

3.3. Term Weighting

Word weighting using the TF-IDF (Term Frequency–Inverse Document Frequency) method aims to determine the importance of a word within a document relative to the entire dataset. This method calculates term frequency while assigning higher weights to words that are unique or rarely appear in other documents (inverse document frequency). The process strengthens the influence of relevant words and diminishes the impact of frequently occurring but less meaningful words. This technique is commonly implemented using

libraries such as scikit-learn to efficiently compute weight values [34]. The TF-IDF results are presented in Table 11.

Table 11. TF-IDF Result

No	Text	TF-IDF
1	SUSAH NGOMONG SAMA ORANG YANG GA TAU DIRI	0.064145
2	macam mw balas tweet klau sy x pandai english	0.064282
3	knp ka?? aku jg gk denger siaran trakhir kk lg. sedih. Mksih ya ka udh jd tmn aku di stiap kk lg siaran	0.091421
4	hujan lbat + petirr aku takut pengen off aja	0.064145
5	Yang tadinya tenang dirumah	0.045280

3.4. Dataset Splitting

The dataset employed in this study consisted of 1,000 text samples distributed across four emotion categories: happy, sad, angry, and fearful, with 250 samples in each category. To ensure an objective evaluation of model performance, the dataset was divided into two subsets with a 90:10 ratio. A total of 900 samples were allocated for training, while the remaining 100 samples were reserved for testing. The 90:10 ratio was selected to provide a sufficiently large portion of training data, allowing the model to better learn patterns from each emotion category. At the same time, the testing subset enabled effective evaluation on previously unseen data, which is crucial for assessing model generalization [29].

The class distribution within the dataset was deliberately balanced to prevent bias toward any particular emotion category. This balance is important to avoid classification results being overly influenced by a majority class [24]. Additionally, a simple validation method, i.e., train–test split, was employed to maintain efficiency in the modeling process, considering the relatively small dataset size. A similar 90:10 split ratio has been adopted in prior studies, such as [8], which demonstrated that this division offers an effective trade-off between training performance and testing evaluation in text classification tasks.

3.5. Classification Modeling

This study implemented three primary algorithms for emotion detection in text: Logistic Regression, K-Nearest Neighbors (KNN), and Naïve Bayes. To enhance model performance, an ensemble bagging approach was applied, combining predictions from multiple models to produce more stable and accurate outcomes.

a. *Logistic regression*

Logistic Regression was chosen due to its capability to handle high-dimensional text data, such as TF-IDF representations, and its robustness when processing sparse datasets. Logistic Regression has also demonstrated competitive performance in text classification tasks. The baseline evaluation results showed an accuracy of 88%, with precision, recall, and F1-score of 0.89, 0.88, and 0.88, respectively. After hyperparameter tuning with ensemble bagging, performance improved to 89% accuracy, with precision, recall, and F1-score of 0.90, 0.89, and 0.89, respectively.

b. *K-Nearest Neighbors (KNN)*

KNN was used as a comparison to Logistic Regression because of its intuitive nature and its ability to capture non-linear patterns within the data. However, this algorithm is sensitive to noise and often requires additional optimization Baseline results indicated an accuracy of 62%, with precision, recall, and F1-score of 0.64, 0.62, and 0.62, respectively. After optimization through ensemble bagging, accuracy improved significantly to 81%, with precision, recall, and F1-score of 0.83, 0.81, and 0.81, respectively.

c. *Naive Bayes*

Naïve Bayes was selected due to its probabilistic approach, which is well-suited for text classification, particularly with relatively small datasets [31]. The baseline model achieved an accuracy of 80%, with precision, recall, and F1-score of 0.81, 0.80, and 0.80, respectively. After optimization with ensemble bagging, accuracy increased to 81%, with precision, recall, and F1-score of 0.83, 0.81, and 0.81, respectively.

d. Model Comparison

At this stage, an analysis was conducted to evaluate the performance of the models used in this study. The objective was to assess how effectively each algorithm could detect emotions with high accuracy, both under baseline conditions and after optimization with ensemble bagging. Table 12 presents a comparison of the performance results for the three algorithms, including outcomes before and after optimization.

Table 12. Evaluasi Model Result

Methode	Accuracy (%)	Precision	Recall	F1-Score
Logistic regression (LR)	88	0,89	0,88	0,88
Ensemble Bagging (LR)	89	0,90	0,89	0,89
KNN	62	0,64	0,62	0,62
Ensemble Bagging (KNN)	81	0,83	0,81	0,81
Naïve Bayes	80	0,81	0,80	0,80
Ensemble Bagging (NB)	81	0,83	0,81	0,81

The Ensemble Bagging models (KNN and Naïve Bayes) achieved accuracy, precision, recall, and F1-score values of 81%. Logistic Regression initially obtained an accuracy of 88%; however, applying ensemble bagging increased its accuracy to 89%, yielding the best performance among all models. On the other hand, the KNN model without bagging showed the lowest performance with an accuracy of 62%, but after applying ensemble bagging, its performance improved significantly to 81%. These results demonstrate that the ensemble bagging method is effective in enhancing model performance.

3.6. Measurement Value

The performance of the models was evaluated using accuracy, precision, recall, and F1-score to measure their ability to detect emotions in Indonesian texts. These metrics were calculated based on the confusion matrix, which illustrates the relationship between predicted and actual values. According to the evaluation results, Logistic Regression achieved the best performance, both before and after optimization with ensemble bagging. Table 13 and Equations (6)–(9) present the performance calculation of Logistic Regression after optimization, based on the following confusion matrix.

Table 13. Predicted and Actual

	Predicted Positive	Predicted Negative
Actual Positive	89 (TP)	11 (FN)
Actual Negative	11 (FP)	89 (TN)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} = \frac{89+89}{89+89+11+11} = \frac{178}{200} = 0{,}89\ (89\%) \quad (6)$$

$$Precision = \frac{TP}{TP+FP} = \frac{89}{89+11} = \frac{89}{100} = 0{,}90 \quad (7)$$

$$Recall = \frac{TP}{TP+FN} = \frac{89}{89+11} = \frac{89}{89+11} = 0{,}89 \quad (8)$$

$$F1\ score = \frac{2(Precision \cdot Recall)}{Precision+Recall} = 2 \times \frac{0{,}90 \times 0{,}89}{0{,}90+0{,}89} = 0{,}895(89{,}5\%) \quad (9)$$

The results indicate that Logistic Regression with ensemble bagging exhibited excellent performance and emerged as the best-performing model in this study. In addition to the aforementioned metrics, the models were also evaluated using the ROC-AUC (Receiver Operating Characteristic – Area Under the Curve). The ROC-AUC curve measures the ability of a model to distinguish between positive and negative classes across various thresholds. The closer the value is to 1, the better the model's classification ability. The following figures illustrate the ROC-AUC curves for the three models employed in this study: Logistic Regression, KNN, and Naïve Bayes.

Figure 2 shows the ROC-AUC curve of the Logistic Regression model in a multiclass classification setting. Each class achieved a high AUC value, with Class 0 and Class 1 reaching 0.99, Class 2 achieving 0.96, and Class 3 achieving 0.98. AUC values close to 1 indicate that the model has excellent discriminative capability among the emotion classes in the dataset. This curve further reinforces the superiority of Logistic Regression as the best-performing model in this research.

Figure 3 illustrates the ROC (Receiver Operating Characteristic) curve of the K-Nearest Neighbors (KNN) algorithm in multiclass classification. The graph presents the model's performance for each class, as represented by varying AUC values. Class 1 achieved the highest performance with an AUC of 0.97,

followed by Class 0 (AUC = 0.90), Class 3 (AUC = 0.88), and Class 2 (AUC = 0.86). This curve demonstrates the trade-off between the False Positive Rate (FPR) and True Positive Rate (TPR) for each class. To improve clarity and reproducibility, the corresponding AUC values for each emotion class are summarized in table 14.

Table 14. ROC-AUC Values per Class for Each Classification Model

Model	Class 0	Class 1	Class 2	Class 3
Logistic Regression	0.99	0.99	0.96	0.98
K-Nearest Neighbors	0.90	0.97	0.86	0.88
Naïve Bayes Classifier	0.92	0.95	0.89	0.91

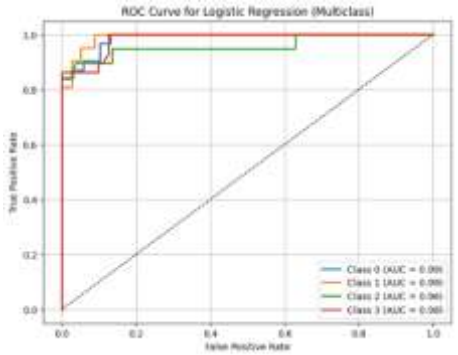

Fig. 2. ROC-AUC Curve for the Logistic Regression Model

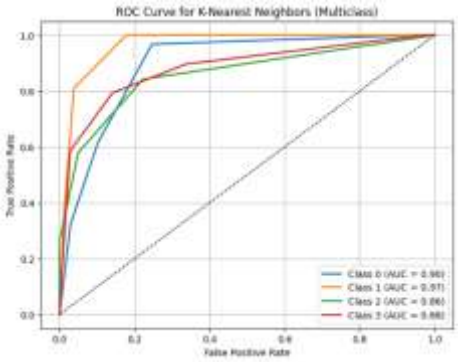

Fig. 3. ROC-AUC Curve for the K-Nearest Neighbors (KNN) Model

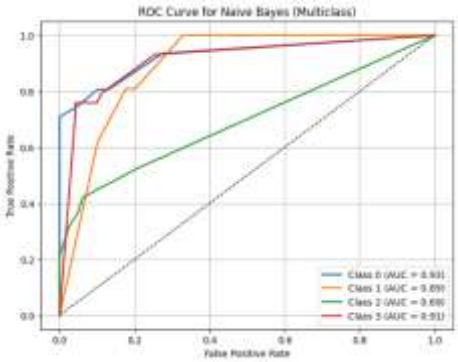

Fig. 4. ROC-AUC Curve for the Naïve Bayes Model

4. Conclusion

This study demonstrates that Logistic Regression optimized with the ensemble bagging approach is an effective solution for detecting emotions in Indonesian texts. The model successfully addressed challenges such as slang variations, the presence of stopwords, and the limited size of Indonesian-language datasets. With an accuracy of 89%, this method achieved the best performance compared to other models such as KNN and Naïve Bayes. Furthermore, the comprehensive preprocessing steps contributed significantly to improving data quality and predictive outcomes. The findings of this study are expected to serve as a reference for the development of Indonesian text-based emotion analysis applications, such as social media analytics or human–computer interaction. However, the limitations of this research lie in the relatively small dataset size and the restricted coverage of emotion categories. Future studies are recommended to employ larger and more diverse datasets to further enhance model generalization..

Acknowledgment

The authors would like to express their gratitude to the Faculty of Engineering, Universitas Suryakancana, for providing a platform to conduct and develop this research. It is hoped that this study will contribute significantly to the advancement of scientific knowledge in Indonesia.

Declarations

Author contribution. ll authors contributed substantially to this research. Contributions include problem formulation, literature review, methodology design, data analysis, and manuscript preparation. All authors have read and approved the final manuscript for publication.
Funding statement. This research was self-funded by the authors.
Conflict of interest. The authors declare no conflict of interest regarding the research or the publication of this article.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The data and processes implemented using Python that support the findings of this study are accessible at the following public repository: [https://github.com/erfianjunianto/deteksi-emosi-mji]. All datasets analyzed or generated during this research/experiment are openly available for academic purposes and further research..

References

[1] A. N. Rohman, E. Utami, and S. Raharjo, "Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing," *Eksplora Inform.*, vol. 9, no. 1, pp. 70–76, Sep. 2019.

[2] J. Bata, Suyoto, and Pranowo, "Leksikon untuk Deteksi dari Teks Bahsa Indonesia," 2018.

[3] R. Sutoyo, H. L. H. S. Warnars, S. M. Isa, and W. Budiharto, "Indonesian Twitter Emotion Recognition Model using Feature Engineering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, pp. 1057–1065, 2023.

[4] F. M. Alotaibi, "Classifying Text-Based Emotions Using Logistic Regression," *VAWKUM Trans. Comput. Sci.*, vol. 7, no. 1, pp. 31–37, Apr. 2019.

[5] D. A. Nur'Faradila, L. Magdalena, and M. Febima, "Penerapan Naïve Bayes Classifier Untuk Klasifikasi Postingan Berita Hoaks Di Instagram Cirebon Saber Hoaks," *Media J. Inform.*, vol. 16, no. 2, p. 243, Dec. 2024.

[6] M. D. R. Wahyudi, "Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur'an Berdasarkan Keterkaitan Topik," *Semesta Tek.*, vol. 22, no. 1, 2019.

[7] F. Rahutomo, A. Retno, and T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 1, pp. 41–48, Jan. 2019.

[8] E. Junianto and R. Rachman, "Penerapan Metode Naïve Bayes Classifier Untuk Mendeteksi Emosi Pada Komentar Media Sosial," *J. Responsif Ris. Sains dan Inform.*, vol. 2, no. 1, pp. 1–8, Feb. 2020.

[9] F. Fanesya, R. C. Wihandika, and I. Indriati, "Deteksi Emosi Pada Twitter Menggunakan Metode Naive Bayes Dan Kombinasi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 7, pp. 6678–6686, Aug. 2019.

[10] APJII, "Asosiasi Penyelenggara Jasa Internet Indonesia - Survei," 2024. [Online]. Available: https://survei.apjii.or.id/survei. [Accessed: 13-Apr-2025].

[11] A. Kurniasih, A. K. Santoso, B. D. Wicaksono, and H. F. Pardede, "Evaluations of Emotion Analysis of Tweets using Bidirectional Long Short Term Memory and Conventional Machine Learning," *J. Teknol. dan Sist. Komput.*, vol. 10, no. 2, Apr. 2022.

[12] A. Glenn, P. LaCasse, and B. Cox, "Emotion classification of Indonesian Tweets using Bidirectional

LSTM," *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9567–9578, May 2023.
[13] R. N. Sofia and D. Supriyadi, "Komparasi Metode Machine Learning dan Deep Learning untuk Deteksi Emosi pada Text di Sosial Media," *JUPITER J. Penelit. Ilmu dan Teknol. Komput.*, vol. 13, no. 2, pp. 130–139, Oct. 2021.
[14] A. B. P. Negara, H. Muhardi, and F. Sajid, "Perbandingan Algoritma Klasifikasi terhadap Emosi Tweet Berbahasa Indonesia," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 2, p. 242, Aug. 2021.
[15] M. Cindo, D. P. Rini, and E. Ermatita, "Literatur Review: Metode Klasifikasi Pada Sentimen Analisis," *Semin. Nas. Teknol. Komput. Sains*, vol. 1, no. 1, pp. 66–70, Feb. 2019.
[16] O. Ndama, I. Bensassi, and E. M. En-Naimi, "The impact of BERT-infused deep learning models on sentiment analysis accuracy in financial news," *Bull. Electr. Eng. Informatics*, vol. 14, no. 2, pp. 1231–1240, Apr. 2025.
[17] A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text," *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 2943–2961, 2022.
[18] K. Machová, M. Szabóova, J. Paralič, and J. Mičko, "Detection of emotion by text analysis using machine learning," *Front. Psychol.*, vol. 14, 2023.
[19] R. Afrinanda, … L. E.-M. J., and undefined 2023, "Hybrid Model for Sentiment Analysis of Bitcoin Prices using Deep Learning Algorithm," *journal.universitasbumigora.ac.id*.
[20] Nanira Annisa Fitri, Taufik Edy Sutanto, and Muhaza Liebenlito, "Deteksi Kepribadian MBTI pada Diskusi Agama Islam di Twitter Indonesia 2009-2019," *Indones. J. Comput. Sci.*, vol. 12, no. 5, Oct. 2023.
[21] E. Junianto, M. Puspitasari, S. I. Zakaria, T. Arifin, I. Wiseto, and P. Agung, "Klasifikasi Emosi pada Teks Berbahasa Inggris Menggunakan Pendekatan Ensemble Bagging," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 13, no. 4, pp. 272–281, Nov. 2024.
[22] A. H. L. Noer and Dedi Wijayanti, "Directive Acts of Speech in Kajian Malam Ahad Kanal Dr. Zaidul Akbar and its Relation to Teaching Materials for Persuasive Texts in Class VIII Junior High Schools," *Aksis J. Pendidik. Bhs. dan Sastra Indones.*, vol. 8, no. 1, pp. 33–47, Jun. 2024.
[23] W. A. Setyati, S. Sunaryo, A. Rezagama, A. K. Widodo, and M. F. A. Yulianto, "Penerapan Regresi Logistik Dalam Penentuan Faktor Yang Mempengaruhi Jumlah Wisatawan Ecotourism Desa Bedono," *J. ENGGANO*, vol. 5, no. 1, pp. 11–22, Apr. 2020.
[24] W. O. Simanjuntak, A. Bijaksana, P. Negara, and R. Septriana, "Perbandingan Algoritma Logistic Regression dan Random Foret (Studi Kasus : Klasifikasi Emosi Tweet)," *J. Apl. dan Ris. Inform.*, vol. 1, no. 2, pp. 160–164, Jun. 2023.
[25] F. Sukmanisa, Y. A. Sari, and I. Cholissodin, "Deteksi Emosi pada Tweet Berbahasa Indonesia tentang Pembelajaran Jarak Jauh Menggunakan K-Nearest Neighbor dengan Pembobotan Kata Term Frequency-Inverse Gravity Moment," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 9, pp. 4033–4041, Sep. 2021.
[26] D. R. Wulandari, C. Setianingsih, and F. M. Dirgantara, "Deteksi Emosi Berbasis Teks Untuk Menganalisis Kuliah Daring Selama Masa Pandemi Menggunakan Algoritme Naive Bayes," *eProceedings Eng.*, vol. 9, no. 4, Aug. 2022.
[27] F. H. Rachman and I. Imamah, "Pendekatan Data Science untuk Mengukur Empati Masyarakat terhadap Pandemi Menggunakan Analisis Sentimen dan Seleksi Fitur," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 492, Dec. 2022.
[28] L. Annisa and A. D. Kalifia, "Analisis Teknik TF-IDF Dalam Identifikasi Faktor-Faktor Penyebab Depresi Pada Individu," *Gudang J. Multidisiplin Ilmu*, vol. 2, no. 1, pp. 302–307, Jan. 2024.
[29] W. Widayat, "Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning," *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 3, p. 1018, Jul. 2021.
[30] D. Onita, "Active Learning Based on Transfer Learning Techniques for Text Classification," *IEEE Access*, vol. 11, pp. 28751–28761, 2023.
[31] F. Rustam, A. Mehmood, M. Ahmad, S. Ullah, D. M. Khan, and G. S. Choi, "Classification of Shopify App User Reviews Using Novel Multi Text Features," *IEEE Access*, vol. 8, pp. 30234–30244, 2020.
[32] N. Ramadhani and N. Fajarianto, "Sistem Informasi Evaluasi Perkuliahan dengan Sentimen Analisis Menggunakan Naïve Bayes dan Smoothing Laplace," *J. Sist. Inf. BISNIS*, vol. 10, no. 2, pp. 228–234, Dec. 2020.
[33] D. Kurniawan, H. D. Purnomo, and A. Iriani, "Analisis Sentimen Komentar Konsumen Industri Jamu di Media Sosial menggunakan Artificial Neural Network dan K-Nearest Neighbor," *J. Sist. Inf. Bisnis*, vol. 14, no. 3, pp. 210–223, Aug. 2024.
[34] F. D. U. Arif, "Perbandingan kinerja algoritma random forest, xgboost dan lightgbm dalam klasifikasi emosi komentar reddit," Fakultas Sains dan Teknologi UIN Syarif HIdayatullah Jakarta, Jakarta, 2024.