# Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences

Amelia Devi Putri Ariyanto[1*], Fari Katul Fikriah[2], and Arif Fitra Setyawan[3]
[1−3]Information Systems and Technology Study Program, Widya Husada University
Semarang, Indonesia 50146
Email: [1]ameliadev26@gmail.com, [2]farichatulfikriyah45@gmail.com, [3]ariffitra.setyawan@gmail.com



***Abstract*—The ability to detect emotions in short texts is crucial because interpreting emotions on platforms like Twitter can offer insight into social trends and responses to specific events. Additionally, examining emotions in product reviews assists companies in comprehending customer sentiment, allowing them to improve the quality of their products and services. Most research on Indonesian language emotion detection utilizes statistical feature extraction, with limited discussion on the impact of both statistical and semantic feature extraction. Thus, the research aims to detect emotions in short texts equipped with an analysis of the impact of statistical and semantic features. Analysis of the impact of statistical and semantic features on short texts is necessary to identify the most effective approaches, improve detection accuracy, and ensure that the developed systems can better handle the variety and complexity of informal language. The data used are a public dataset originating from Twitter texts and product review texts in e-commerce. The research utilizes statistical features such as Term Frequency Inverse Document Frequency (TF-IDF) and semantic features such as Bidirectional Encoder Representations from Transformers (BERT). The evaluation results show that using semantic features significantly improves the performance of emotion detection in short texts by 13–24%. It is higher than using statistical features. Deep Learning (DL) algorithms based on neural networks have also been proven to outperform Machine Learning (ML) algorithms in detecting emotions in short text. The experimental results and outlines show the potential directions for future development.**

***Index Terms*—Emotion Detection, Semantic Features, Statistical Features, Machine Learning, Short Texts**

## I. Introduction

EMOTIONS encompass an individual's mental and physical state connected to thoughts, feelings, and behavioral reactions, which can be recognized through speech and facial expressions [1]. Emotion detection attempts to capture emotional nuances from data, which can be in sound, images, or text [2]. Shaver's theory classifies emotions into five basic categories: happiness, love, sadness, anger, and fear [3]. The five emotional categories in the text data have certain lexicon words, such as the words "bad", "hate", and "suspicious", which tend to be associated with the emotion of anger [4].

Social media becomes a new trend for interaction and communication, so the number of social media users continues to increase, especially on Twitter [5]. Indonesia ranks third in the number of active Twitter users in Asia Pacific [1], so making posts on Twitter (or referred to as tweets) has great potential to be used in emotion detection. Tweets are often posted in real-time and provide insight into ongoing emotional trends in society. As a result, public opinion on various topics, such as politics or social issues, can be effectively monitored via Twitter (which has changed its name to X). When analyzing emotions in informal text data, it is crucial to consider various sources, such as social media discussions covering general topics and consumer reviews on e-commerce platforms that specifically focus on the experience of using products.

The rapid progress in e-commerce has also changed people's way of shopping to become completely online because it can provide convenience and the offered prices are more affordable for consumers. However, there is a mismatch between the actual quality of the product and the description provided by the sellers on the e-commerce platform, resulting in many consumers looking for product information through e-commerce reviews that cover various aspects such as price or service [6]. Automatic emotion detection is important in product review texts in e-commerce because it can help companies to understand customers' feelings and attitudes toward the products or services being sold

Received: May 30, 2024; received in revised form: Oct. 22, 2024; accepted: Oct. 23, 2024; available online: Apr. 14, 2025.
*Corresponding Author

Cite this article as: A. D. P. Ariyanto, F. K. Fikriah, and A. F. Setyawan, "Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences", CommIT Journal 19(1), 1–13, 2025.

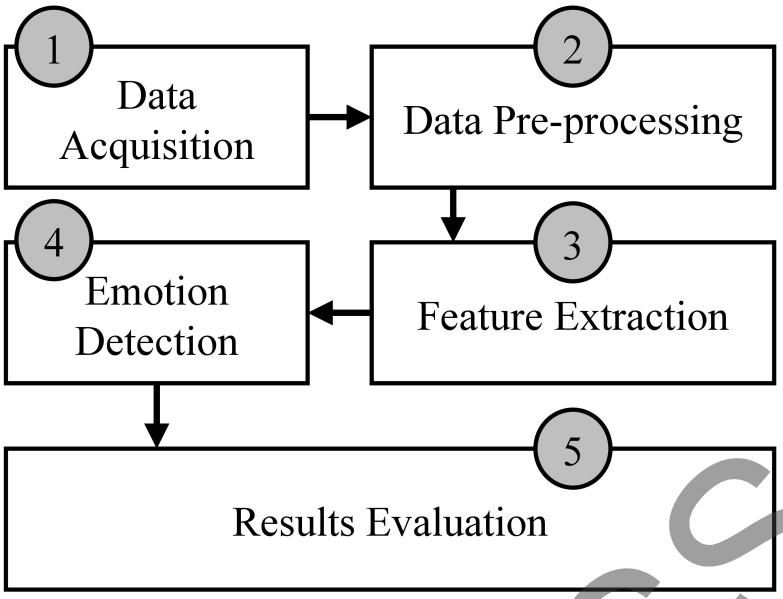


Fig. 1. General description of text-based emotion detection.

based on their reviews [7]. Automatic emotion detection can aid governments, nonprofits, and companies to create more effective policies and communication strategies [8].

Previous research states five main stages for extracting emotions from the text (Fig. 1): data acquisition, data pre-processing, feature extraction, emotion detection, and results evaluation [2]. Data acquisition is the process of collecting data for emotion detection. Text data for emotion detection in Indonesian can be formal or informal. The characteristics of formal text are an organized sentence structure, standard language, and no use of abbreviations [9]. Then, another previous research uses formal Indonesian texts from online folk tales [10]. Detecting emotions in short sentences, such as those found on Twitter or in e-commerce product reviews, presents unique challenges compared with formal text. Daily human communication creates texts with distinctive features such as short sentences, copious slang, irregular word structures, and abundant abbreviations [11]. Detecting emotions from short text sentences requires additional processing. Several public datasets are available for detecting Indonesian language emotions in informal texts. For example, dataset developed by [1] uses text data from Twitter, and [4] uses product review text data from e-commerce.

Next, data pre-processing is performed after the data acquisition process [2]. Data pre-processing transforms raw text data into a structured format [14]. The output of the data preprocessing process is clean text data. Then, feature extraction must also be performed before being entered into the Machine Learning (ML) algorithm to determine the emotion. Feature extraction is a method for representing text documents into numerical vectors because ML algorithms generally require input in numerical form [15]. Numerical vector representation from the text data feature extraction process can produce more compact and easy-to-manage features, even though raw text data can contain thousands or millions of words.

Previous Indonesian text emotion detection research primarily employs statistical feature extraction, specifically, Term Frequency Inverse Document Frequency (TF-IDF). Previous research [5] analyzes the public emotionally via Twitter regarding government policies during the COVID-19 pandemic. The feature extraction carried out in the research is TF-IDF, which outperforms other feature extraction methods such as FastText. Previous research [12] also detects emotions in Indonesian on Twitter text data using TF-IDF statistical features. Another previous research [10] identifies emotions in Indonesian formal texts such as online folk tales. TF-IDF converts fairy tale texts into numerical vectors via ML algorithms before emotion detection. Then, another previous research [13] uses TF-IDF feature extraction for emotion detection in Indonesian Twitter text and finds that statistical feature extraction (TF-IDF) performs better than lexicon-based feature extraction. TF-IDF concentrates on determining the frequency of words in a text document [16] but fails to encapsulate the intricate meanings conveyed through the usage of words, particularly in emotion detection

Cite this article as: A. D. P. Ariyanto, F. K. Fikriah, and A. F. Setyawan, "Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences", CommIT Journal 19(1), 1–13, 2025.

TABLE I
DIFFERENCES IN CONTRIBUTION WITH SEVERAL PREVIOUS RESEARCH REGARDING EMOTION DETECTION IN INDONESIAN TEXTS.

| Author and Year of Publication | Previous Research | Current Research |
|---|---|---|
| [5], 2023 | Detection of people's emotions about government policies via Twitter is conducted.<br>It does not perform statistical or semantic feature analysis. | The research analyzes statistical and semantic feature extraction for emotion detection in informal Indonesian language texts, including Twitter posts and e-commerce product reviews.<br>The algorithm uses for Deep Learning (DL)-based emotion detection (Multilayer Perceptron (MLP)). |
| [12], 2022 | It compares the performance of several Machine Learning (ML) algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree (DT).<br>Neighbor (KNN), and Decision Tree (DT), using statistical features such as Term Frequency Inverse Document Frequency (TF-IDF). | It conducts statistical and semantic feature extraction analysis for emotion detection in informal Indonesian texts.<br>The algorithm used for emotion detection is Deep Learning (DL) based on neural networks. |
| [10], 2016 | It detects emotions in Indonesians based on formal text data in the form of online folk tales.<br>The statistical feature extraction method (Term Frequency Inverse Document Frequency (TF-IDF)) represents text in a numerical vector.<br>The Machine Learning (ML) algorithm used for emotion detection uses Naïve Bayes (NB), J48, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN).<br>The analysis carried out is only for the performance of the Machine Learning (ML) algorithm. | The research performs Indonesian emotion detection using informal text data from Twitter and e-commerce product reviews.<br>The research analyzes the impact of statistical- and semantic-based feature extraction on informal text because feature extraction is an important process in emotion detection to convert raw text into numerical vectors.<br>The algorithm uses for Deep Learning (DL)-based emotion detection is Multilayer Perceptron (MLP). |
| [13], 2022 | Emotion recognition is performed using Indonesian Twitter data.<br>The research only carries out Term Frequency Inverse Document Frequency (TF-IDF) feature extraction analysis with Lexicon. | The research performs statistical and semantic feature extraction<br>analysis for emotion detection in informal Indonesian texts. |

scenarios.

The development of contextual embedding has led to enhancements in feature extraction for Natural Language Processing (NLP) tasks, including emotion detection and sentiment analysis. Contextual embedding is a vector representation of words, phrases, or texts that consider context. Each word is given a unique vector representation depending on other words in the sentence or document in question [17]. Considering the context enhances the accuracy of word meanings in a sentence. The Bidirectional Encoder Representations from Transformers (BERT) model, which utilizes transformer-based approaches to process text bidirectionally, considering context from previous and succeeding words, is a widely-used contextual embedding method [18].

When the text data is represented as a numerical vector, it can be processed by an ML algorithm for emotion detection. Several previous studies [5, 10, 12, 13] mostly use ML algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes (NB), or Random Forest (RF). However, not many studies use neural network-based Deep Learning (DL) algorithms, such as the Multilayer Perceptron (MLP) algorithm in Indonesian language emotion detection. MLP can effectively model nonlinear relationships between text features and emotional labels in complex, nuanced, and informal Twitter text data.

The research adopts five main emotion detection stages, as described [2]. Most research on Indonesian language emotion detection utilizes statistical feature extraction, with few studies examining the impact of both statistical and semantic feature extraction. Informal short texts convey imply meanings or emotions indirectly, which is necessary to analyze the impact of semantic features and detect emotions that can help to capture the implied meaning of short texts. The differences between the contributions of the research and those of previous studies are summarized in Table I.

Unlike previous studies, the main objective of the research is to explicitly analyze the impact of statistical and semantic feature extraction for emotion detection in short Indonesian text sentences. A comparative research is conducted to answer several research questions: 1) how is the performance of statistical and semantic feature extraction for emotion detection in short Indonesian text sentences? 2) Can semantic feature extraction improve performance? 3) How is the performance of ML and DL algorithms in detecting emotions? The research results can pave the way for researchers to develop more sophisticated analysis techniques for emotion detection in short texts, such as the exploration of multi-modal emotion analysis or recognition of emotional expressions in various Indonesian dialects. Moreover, feature extraction is crucial in NLP tasks because ML/DL algorithms can only process numeric inputs. Short informal text sentences with

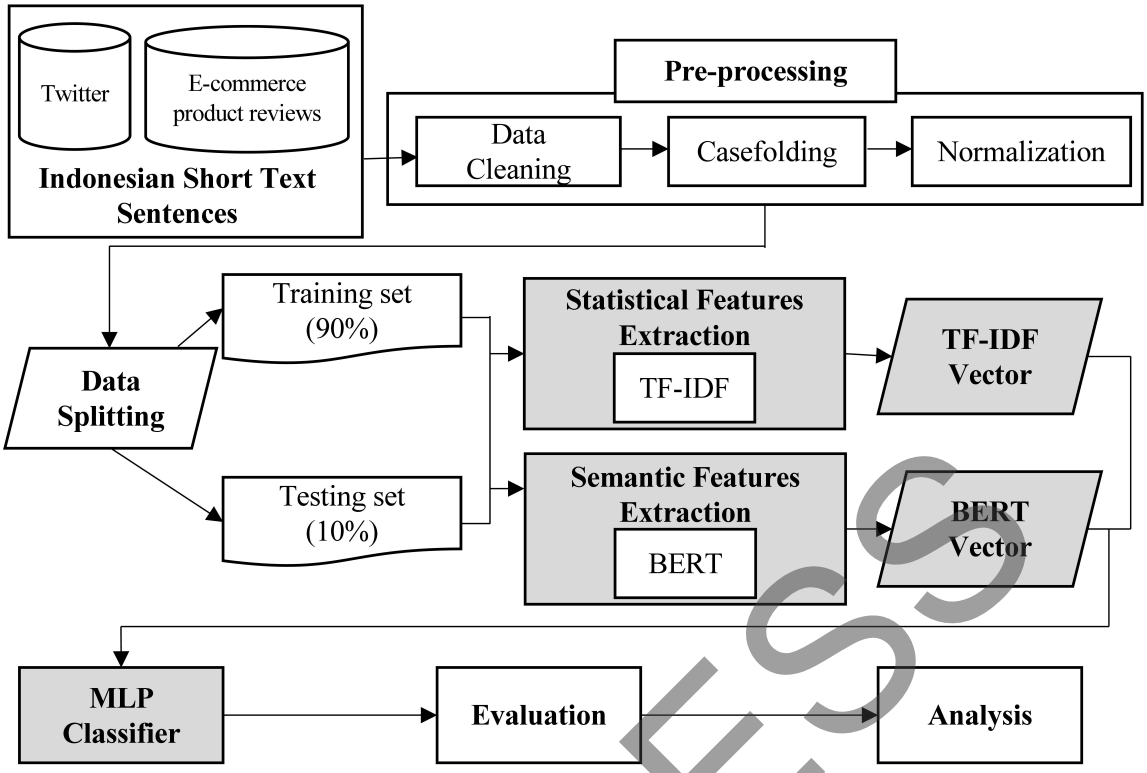


Fig. 2. Flow diagram of emotion detection in Indonesian short texts using statistical and semantic features with the Multilayer Perceptron (MLP) algorithm. It has Term Frequency Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT)



unique characteristics require in-depth study, so feature analysis is performed using a neural network-based DL algorithm, such as MLP, for emotion detection in Indonesian short texts.

## II. Research Method

The research adopts the five main stages for emotion detection described by previous research [2]. It consists of (1) data acquisition, (2) data pre-processing, (3) feature extraction, (4) emotion detection, and (5) results evaluation. The researchers analyze the impact of statistical and semantic extraction features on short informal texts detecting emotions during the third stage. Emotion detection relies on a neural network-based DL method, specifically MLP. Figure 2 shows the flow of the research method. The details of each stage are explained in the following subsection.

### A. Data Acquisition

The data used in the research are a public dataset originating from Twitter texts [1] and product review texts in e-commerce [4]. The review text is taken from one of the largest e-commerce sites in Indonesia, Tokopedia. Some product categories reviewed include kids and baby fashion, electronics, beauty and body care, and several other products. The total number of product reviews in the public dataset reached 5,400 review texts, with the distribution of each emotion label for love, happiness, anger, sadness, and fear being 809, 1770, 699, 920, and 1202 [4], respectively. Meanwhile, the Indonesian language public Twitter dataset ws taken from June 1, 2018, to June 14, 2018, with a distribution of emotion labels of 1101, 1017, 997, 649, and 637 for the emotions of anger, happiness, sadness, fear, and love, respectively [1]. The data are adjusted for each emotion label to have an equal representation of 600 texts in both datasets, disregarding the issue of data imbalance. The characteristics of each emotion label are explained in Table II.

### B. Data Preprocessing

In e-commerce and Twitter communications, texts are written in an informal and everyday conversational style to keep them casual and unstructured. Therefore, preprocessing data is needed to make the two texts cleaner and more structured before being fed into the ML algorithm to detect emotions. The first stage of data pre-processing is data cleaning, which includes re-

TABLE II
DEFINITION OF EACH EMOTION LABEL AND EXAMPLE SENTENCES.

| Emotion Labels | Sentence Characteristics of Each Label | Example Sentences (in E-Commerce) | Example Sentences (on Twitter) |
|---|---|---|---|
| Anger | Expressions that reflect a user's anger, frustration, or displeasure which can be expressed explicitly through strong and negative words. | *Pelayanan toko jelek, proses barang lambat, di chat tidak menjawab, barang rata-rata* (Poor shop service, slow processing of goods, no response to chat, average goods) | *Gemes bgt sama orang yg masih menganggap wajar pelecehan sexual secara verbal mau verbal atau fisik atau apapun, itu tetap tidak dibenarkan!* (I am annoyed with people who still consider verbal sexual harassment normal, whether verbal or physical. It is still not justified!) |
| Fear | Phrases that reflect users' fears, worries, anxieties, or insecurity. | *hati hati ada jebakan*. be careful. *judul dan deskripsi tertulis Dart Game Besar 17 isi paket: 1 buah papan Dart 17 INCH, kenyataannya, yg dikirim hanya ukuran 15 (15 INCH)*. be careful. ((Be careful, there are traps. be careful. The title and description say Dart Game Large 17. Package contents: 1 Dartboard 17 INCH. Only size 15 (15 INCH) was sent. be careful.) | *aku sama kayak kamu btw trauma sama laki-laki karna ayah. semenjak lulus sekolah aku puber suka cowo lah, cowo deketin aku lah segala macem. aku kira trauma ku sembuh tapi berkali-kali aku dikecewain laki-laki* (I'm just like you, btw. I am traumatized by men because of my father. Since graduating from school, I have hit puberty. I like guys, guys come close to me, all kinds of things. I thought my trauma had healed, but many times, I was disappointed by men) |
| Happiness | Expressions that reflect the user's happiness, satisfaction, or joy can be expressed explicitly through positive words such as praise, appreciation, and enthusiasm. | *Barang bagus, pengemasan aman, dapat berfungsi dengan baik* (Good item, safe packaging, works well) | *alhamdulillah, sekarang kurma bukan hanya bisa di panen di tanah* Arab Saudi *tetapi juga bisa di panen di* Pekanbaru (Riau) (Thank God, now dates can be harvested not only in Saudi Arabia but also in Pekanbaru (Riau)) |
| Love | Expressions that reflect a user's feelings of love, warmth, admiration, or deep affection for a product or service. | *Cakep produk nya, rekomen banget nih, sesuai deskripsi* (The product is beautiful and I highly recommend it, according to the description) | *gue sayang mereka ya allah*. i want them to be happy in their entire life, *pen mereka sukses, dapet jodoh yang baik, punya rumah tangga yang langgeng, punya anak yg lucu, gue selalu doa semoga kita ttp barengan sampe udah gede, jd ibuibu, sampe nenek nenek*. (I love them, oh! my God. I want them to be happy in their entire life, be successful, find a suitable mate, have a long-lasting household, and have cute children. I always pray that we will stay together until we are older and become mothers and grandparents.) |
| Sadness | Phrases that reflect the user's sadness, disappointment, loneliness, or other negative feelings. These negative feelings can take the form of feelings of dissatisfaction with certain products or services or complaints about bad experiences | *Kecewa parah nggk berkah jualan gitu* (very disappointed with no luck selling like that) | *kenapa sekarang bisa beda banget sama pas sma ya di tanah rantau bener bener sadar kalo gak semua orang itu bahagia,gak semua orang itu tulus* (Why is it so different now from when I was in high school in overseas countries? I realize that not everyone is happy and not everyone is sincere.) |



moving links, extra spaces, hashtags, the word retweet, which is usually written as "RT" and punctuation.

Next, case folding is performed by changing all text to lowercase. Eliminating the difference between upper and lower case letters can reduce the number of word variations that must be processed by ML algorithms and increase the efficiency of emotion detection tasks. The final data preprocessing stage is normalization, which converts nonstandard text into a formal form [19]. The previous research uses a colloquial lexicon [20] to normalize informal texts because review texts in e-commerce and texts on Twitter contain many spelling errors, abbreviations, slang, and nonstandard language styles. Normalization of an informal text reduces misspellings, abbreviations, and slang words, making it more straightforward for ML algorithms to analyze and improve their performance.

After cleaning the text data, the researchers split it into training and testing sets. Data splitting is useful for evaluating how well a model can generalize the patterns it has learned from training data to data it has never seen before (testing data). The researchers allocate 90% of the data for training and the remaining 10% for testing, following the methodology of [21]. The model's pattern and feature detection capabilities are refined using 90% of the training data.

### C. Feature Extraction

The research uses statistical (TF-IDF) and semantic (BERT) features for emotion detection in short informal texts and assesses their respective impacts. TF-IDF assigns weights to words based on their frequency and rarity across documents, enabling the identification of significant terms relevant to emotion expression. Meanwhile, BERT captures deep contextual representations by leveraging bidirectional encoding, producing

meaningful vector embeddings that enhance text understanding. The research adopts a feature-based approach with IndoBERT, averaging token embeddings from the second-to-last hidden layer to extract semantic features efficiently. These extracted features are then fed into an MLP model chosen for its ability to learn complex patterns and non-linear relationships in textual data.

*1) Statistical Feature:* TF-IDF gives higher weight to words that appear frequently but rarely in a document, so it can help to identify the most relevant words in a particular context. The calculation of Term Frequency (TF) in the first stage of TF-IDF involves measuring the frequency of words in a document, with more frequent words receiving higher TF values. The formula for calculating TF is shown in Eq. (1), where $W_{TF}$ is the TF weight calculated using the logarithm function. Then, $t$ represents a specific term (word) within the text, and $d$ denotes a particular document in the corpus. The TF measures how often the term $t$ appears in document $d$, helping to determine the relative importance of that term within the document.

$$W_{TF_{t,d}} = \begin{cases} 1 + \log_{10}(TF_{t,d}), & \text{if } TF_{t,d} > 0 \\ 0, & \text{if } TF_{t,d} < 0 \end{cases}. \quad (1)$$

Next, the researchers calculate the Inverse Document Frequency (IDF) to measure word rarity in the entire document collection. The IDF formula is given by Eq. (2). The logarithm of the total number of documents is divided by the number of documents containing that word. It has $N$ as the number of text documents and $DF_t$ or document frequency as the number of documents containing the word $t$. Words frequently appearing in many documents will have a lower IDF value, whereas words that rarely appear will have a higher IDF value [16].

$$W_{IDF_t} = \log_{10}\left(\frac{N}{DF_t}\right). \quad (2)$$

Next, the researchers multiply TF and IDF to obtain the final score. The result of this multiplication is a weight for each word in the document that reflects the importance of that word in a particular document relative to the entire collection of documents, as formulated in Eq. (3). The $W_{t,d}$ is the TF-IDF word weighting. The results of word weighting using TF-IDF form feature vectors can be used in ML models for various applications, including emotion detection.

$$W_{t,d} = W_{TF(t,d)} \times W_{IDF_t}. \quad (3)$$

*2) Semantic Feature:* Semantic features are obtained using BERT, which uses 12 encode blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence lengths, and approximately 110 M total parameters. The strategy for using BERT is divided into the fine-tuning approach and the feature-based approach [22]. The fine-tuning approach starts by training a BERT model on a general language understanding task using a large corpus, then adding one or more classification layers to the BERT model for a specific task, such as emotion detection. The feature-based approach uses a BERT model trained to process the text and produce a vector representation. The representation is taken from the last layer or the average of several layers. The vector representation produced by BERT is stored as a feature representing text. Next, a separate ML model (such as a neural network) is trained using the feature representations generated by BERT.

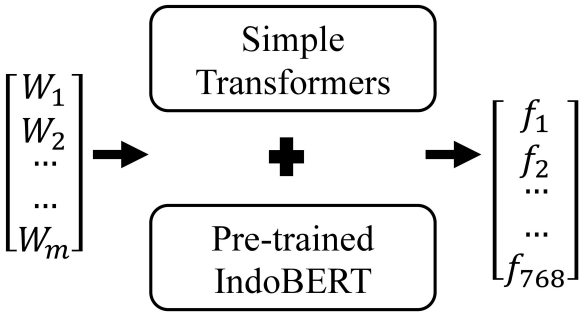


Fig. 3. Feature-based approach for semantic features from pre-trained IndoBERT.

The research employs a feature-based strategy with BERT's robust feature representation for computational and time efficiency, as shown in Fig. 3. The $W$ represents individual words (tokens) from the input text, which are then processed using Pretrained IndoBERT + Simple Transformer to generate corresponding feature representations. The resulting $f$ denotes the extracted feature vectors for each word, capturing contextual semantics in a high-dimensional space.

The Simple Transformers library is applied in research to extract activations from one or more layers without fine-tuning parameters in the original BERT model. The research employs IndoBERT, a pre-trained BERT variant from hugging face that has been trained on informal Indonesian language data using the base configuration. Then, the average pooling for the second-to-last hidden layer of all tokens is calculated to obtain semantic features with a vector size of 768 features. The resulting representation is input for the ML algorithm for emotion detection.

### *D. Emotion Detection*

The research uses a neural network-based DL algorithm, namely MLP, because of its ability to learn complex and non-linear patterns from text data on

Twitter and e-commerce. MLP has a hidden layer that can transform input data into a more complex representation, where each layer performs a non-linear transformation using an activation function. With weights and biases that can be updated through backpropagation, MLP can also find the optimal representation to separate existing classes. Unlike traditional ML algorithms such as NB, it assumes that each feature is independent and only calculates the probability of word occurrence based on frequency so that it cannot capture interactions between words [18]. Other traditional ML algorithms, such as Logistic Regression (LR), can only create linear decision boundaries. It limits usage to datasets that can be separated linearly. The research selects the MLP algorithm for emotion detection because of its advantages.

The basic structure of MLP is input, hidden, and output layers [23], as illustrated in Fig. 4. First, data are entered into the input layers. Then, each neuron in the hidden layer receives the input and processes it with a nonlinear activation function. The purpose of the non-linear activation function ($h$) is to change the input from the previous layer to introduce complexity and the ability of the model to capture non-linear relationships in the data. Next, the output of one hidden layer $a^{k-1}$ becomes the input for the next hidden layer $a^k$, and so on until the output layer, which is formulated in Eq. (4). The MLP model also attempts to find the optimal weight that minimizes the prediction error (Eq. (5)) by finding the optimal weight vector ($B^*$) and minimizing the weight ($B$) of the total loss of each data instance. The loss function ($L$) is chosen based on the basis of the task at hand, such as a loss function for classification. Symbol $y_n$ is the actual label of the data, while $f(x_n; B)$ is the prediction made by the model.

$$a^k = h\left(B^k a^{k-1}\right), \tag{4}$$

$$B^* = \min_B \sum_n L(y_n, f(x_n; B)). \tag{5}$$

The research uses the Scikit-learn Library while implementing the MLP algorithm. The hidden layer used is (100, 5), where the first hidden layer consists of 100 neurons to capture many features and patterns from the input data. The second hidden layer consists of five neurons that act as filters. They filter important information learned by the first layer and reduce complexity before it reaches the output layer. Multiple hidden layers of different sizes can help the network to learn more complex and detailed representations.

Next, the activation function used is Rectified Linear Unit (ReLU), which is defined in Eq. (6). The $f(x)$ represents the activation function's output, determining a neuron's activated value. The variable $x$ is the input

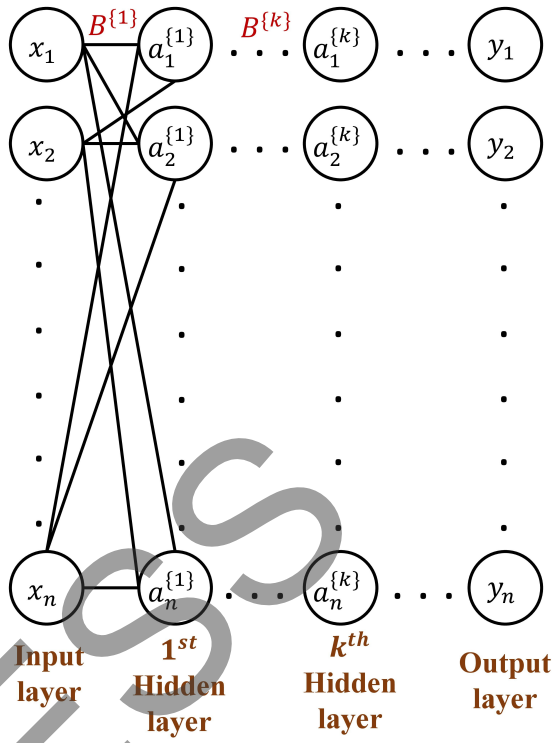


Fig. 4. MLP algorithm structure consists of the input layer ($x_1$ to $x_n$), hidden layer ($a^1$ to $a^k$), and output layer ($y_1$ to $y_n$).

to the function, resulting from a linear transformation applied to the previous layer's output. It is typically computed as a weighted sum of inputs plus a bias term.

$$f(x) = \max(0, x). \tag{6}$$

Vanishing gradient is a problem that arises when training neural networks, where a slight gradient causes insignificant weight updates. The gradient is the derivative of the loss function over the parameters. As a result, the training becomes very slow or even stops altogether. ReLU is chosen because it can overcome vanishing gradients. After all, the output is zero for every negative input, and for every positive input, the output is the input itself. Weight adjustments remain substantial when the gradient stays constant at 1 for positive values, ensuring effective learning [23].

### *E. Results Evaluation*

The evaluation of model performance for emotion detection can be analyzed by obtaining a confusion matrix. The evaluation table shows the number of correct and incorrect predictions for each class [24]. The emotion detection labels in the research consist of five emotion labels (love, happiness, anger, sadness, and fear) as a multiclass classification. The confusion

matrix for multiclass classification is an extension of the confusion matrix for binary classification. If there are $n$ classes, the confusion matrix will be a matrix of size $n \times n$. Rows represent actual labels, and columns represent predicted labels. Each cell in the matrix $(i, j)$ indicates the number of examples whose actual class is $i$ and is predicted as $j$ by the model.

Based on the confusion matrix, precision, recall, F1-score, and accuracy can be calculated for each class and overall using Eq. (7)–(10), respectively. True Positive (TP) is the number of cases in which the model correctly predicted the positive class. False Positive (FP) is the number of cases where the model incorrectly predicts the positive class (prediction is positive but the actual result is negative). In contrast, False Negative (FN) indicates the number of cases where the model incorrectly predicts the negative class (the prediction is negative but the result is positive).

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{7}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{8}$$

$$\text{F1-Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \tag{9}$$

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} TP_i}{\text{Total Samples}}. \tag{10}$$

## III. Results and Discussion

Emotion detection is performed using two datasets, informal Indonesian texts on Twitter and reviews on e-commerce, with five emotion labels. Neural network-based DL algorithms such as MLP are used to train and test data proportions of 90% and 10%. To analyze the impact of statistical and semantic feature extraction for emotion detection in short Indonesian text sentences, the researchers examine their performance differences, as shown in Table III. The F1-score for semantic feature extraction is 0.72, higher than the 0.58 F1-score for statistical feature extraction when using the same Twitter dataset. F1-score improves by 24.14% according to Eq. (11) ($X$ is the new value, and $Y$ is the old one). The same thing also happens in the review dataset in e-commerce, where the F1-score obtained with semantic feature extraction (0.60) is superior to the F1-score obtained with statistical feature extraction (0.53), with a performance increase of 13.21%.

$$\%Increase = \left(\frac{X - Y}{Y}\right) \times 100\%. \tag{11}$$

The results show that semantic feature extraction is better than statistical feature extraction in informal Indonesian text, with a 13–24% performance increase. Obtaining evaluation scores using BERT as semantic features offers rich semantic information by converting words or phrases into meaning-reflective vectors, making model understanding more accessible. In the context of emotion detection, emotions are conveyed through extensive sentences or paragraphs. TF-IDF cannot comprehend sentence context and word relationships using the statistical feature. The reason is that the TF-IDF calculation considers the frequency of occurrence of words independently and the characteristics of informal texts, which tend to be short, so models with statistical features fail to understand polysemy well. Polysemy means that different words have the same meaning in a certain context or that one word can have different meanings depending on the context [22].

Next, the confusion matrix from the MLP algorithm with statistical features on the e-commerce dataset is shown in Table IV, representing the lowest evaluation results. The numbers on the top left to bottom right diagonal show the amount of data correctly classified for each emotion label, whereas the numbers outside the diagonal show the prediction error by the model. Around 32 pieces of data are correctly predicted as happy emotions, but many models misclassify them as love. The same thing happens with love, where 41 pieces of data are correctly predicted as love, but many of the models misclassify them as happy emotions.

Table V details the statistical and semantic mistakes in the model's predictions for short texts. The first example sentence in Table V indicates the context of love and the optimistic anticipation of the beloved's future, making it a representation of love. However, because the word "happy" is used in the first example sentence, a model with statistical features (TF-IDF) incorrectly classifies it as happiness. The second example sentence conveys a sense of contentment and security, as suggested by the terms "*mantap* (steady)" and "*aman* (safe)". In another setting, these words can convey love or appreciation. Statistical models may struggle to distinguish emotions accurately from short texts, such as those on Twitter or e-commerce reviews. The same thing happens in the third example sentence, where the sentence also shows appreciation and possible love for the services provided. However, the positive nature of the review makes it easy for the model to classify happy emotions using statistical features. The inability of statistical features such as TF-IDF to capture nuances and emotional context in short texts often results in misclassifying the emotions of love and happiness. Meanwhile, when using semantic features such as BERT, the three sentences are correctly classified according to the actual label

Cite this article as: A. D. P. Ariyanto, F. K. Fikriah, and A. F. Setyawan, "Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences", CommIT Journal 19(1), 1–13, 2025.

TABLE III
EVALUATION RESULTS OF THE MULTILAYER PERCEPTRON (MLP) ALGORITHM FOR EMOTION DETECTION IN TEXT DATA ON TWITTER AND PRODUCT REVIEW IN E-COMMERCE USING SEMANTIC AND STATISTICAL FEATURES.

| Feature Extraction | Dataset | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|---|
| Semantic (BERT) | E-Commerce | 0.60 | 0.60 | 0.60 | 0.60 |
| | Twitter | 0.72 | 0.72 | 0.72 | 0.72 |
| Statistical (TF-IDF) | E-Commerce | 0.54 | 0.53 | 0.53 | 0.53 |
| | Twitter | 0.63 | 0.56 | 0.58 | 0.56 |

TABLE IV
CONFUSION MATRIX FROM THE MULTILAYER PERCEPTRON (MLP) ALGORITHM FOR EMOTION DETECTION IN E-COMMERCE DATASET USING STATISTICAL FEATURES.

| | | Predicted Labels | | | | |
|---|---|---|---|---|---|---|
| | | Happiness | Sadness | Fear | Love | Anger |
| Actual Labels | Happiness | 32 | 1 | 1 | 12 | 1 |
| | Sadness | 1 | 40 | 16 | 1 | 11 |
| | Fear | 1 | 21 | 25 | 0 | 16 |
| | Love | 16 | 6 | 1 | 41 | 3 |
| | Anger | 1 | 15 | 15 | 1 | 22 |

TABLE V
EXAMPLE OF MISPREDICTED SENTENCES IN EMOTION DETECTION USING TEXT DATA FROM TWITTER AND E-COMMERCE.

| # | Short Text Sentences in Indonesian | Short Text Sentences in English | Semantic Feature (BERT) | | Statistical Feature (TF-IDF) | |
|---|---|---|---|---|---|---|
| | | | Actual Label | Predicted Label | Actual Label | Predicted Label |
| 1 | *gue sayang mereka ya allah.* i want them to be happy in their entire life, *pen mereka sukses, dapet jodoh yang baik, punya rumah tangga yang langgeng, punya anak yg lucu, gue selalu doa semoga kita ttp barengan sampe udah gede, jd ibuibu, sampe nenek nenek.* | I love them, oh! my God. I want them to be happy in their entire life, be successful, find a good mate, have a long-lasting household, and have cute children. I always pray that we stay together until we are older and become mothers and grandparents. | Love | Love | Love | Happiness |
| 2 | *mantap barang nyaa aman alhamdulillah* | great, the item is safe, thank God | Happiness | Happiness | Happiness | Love |
| 3 | *pelayanan bagus besok2* order *lagi* | good service, tomorrow I will order again | Love | Love | Love | Happiness |
| 4 | *ini bener2 bikin tereak balik kanan nangis jejeritan, waktu di rumah mbak di trihanggo, malem2 nyuci lanjut jemur di samping rumah, dan... kayak ada yg siul. padahal gelap. pas diamatin, buset. sugus kain mori nya udah kusem mukaknya hitam* | This makes me scream and cry. When I was at home in Trihanggo, I washed clothes again in the evening and continued to dry them besides the house. and... it's like someone is whistling. even though it's dark. When I looked at it, it was, my gosh. His mori cloth had already faded, and his face was black. | Fear | Fear | Fear | Sadness |
| 5 | *aku sama kayak kamu* btw *trauma sama laki-laki karna ayah. semenjak lulus sekolah aku puber suka cowo lah, cowo deketin aku lah segala macem. aku kira trauma ku sembuh tapi berkali-kali aku dikecewain laki-laki* | I'm just like you, btw. I am traumatized by men because of my father. Since graduating from school, I have hit puberty. I like guys, guys come close to me, all kinds of things. I thought my trauma had healed, but many times I was disappointed by men | Fear | Fear | Fear | Sadness |
| 6 | *besi pengganjal ga ada...jd ga bisa mengunci ...* | There's no iron stop ... so it cannot be locked ... | Sadness | Sadness | Sadness | Fear |
| 7 | *westerima kenyataan mbah, anak lanange lebih kelihatan muda menawan, tegap dan gagah, ketimbang suaminya cemburuuuuu suami merasa anak lanange lebih sering dibanggakan dari pada suamine dewe* | just accept reality, grandma, her son looks younger, charming, sturdy, and dashing, rather than her husband. jealous. husband feels that his son is more prided upon than him | Sadness | Love | Sadness | Sadness |

Note: Bidirectional Encoder Representations from Transformers (BERT) and Term Frequency Inverse Document Frequency (TF-IDF).

TABLE VI
PERFORMANCE COMPARISON OF SEVERAL ALGORITHMS USING SEMANTIC FEATURES ON TEXT DATA FROM TWITTER.

| Model | Evaluation Metrics | | | |
|---|---|---|---|---|
| | Precision | Recall | F1-Score | Accuracy |
| Multilayer Perceptron (MLP) | 0.72 | 0.72 | 0.72 | 0.72 |
| Decision Tree (DT) | 0.56 | 0.56 | 0.56 | 0.56 |
| Naïve Bayes (NB) | 0.63 | 0.63 | 63 | 0.63 |
| K-Nearest Neighbors (KNN) | 0.63 | 0.65 | 0.64 | 0.65 |
| Logistic Regression | 0.65 | 0.66 | 0.65 | 0.66 |

because BERT can consider the context of the words in the sentence more effectively in detecting emotions in informal text.

Based on Table IV, 40 data instances are correctly classified as sadness, but several statistical feature models mistakenly identify them as fear. The same thing happens with the fear. Around 25 data instances are correctly predicted as fear, but many models with statistical features misclassify them as sadness. The main reason for the prediction error between the fear and sadness labels in models with statistical features is that these two emotions often use similar words or have similar negative nuances. Words indicating fear and sadness can overlap in their use, especially in a negative context.

For example, the fourth example sentence, which describes a frightening situation, clearly shows the emotion of fear, as shown in Table V. However, because the word "*nangis* (crying)" is usually associated with sadness. Models with statistical features such as TF-IDF incorrectly classify the emotion as sadness. In short texts, it is not always possible to distinguish varying emotions because of insufficient context. TF-IDF's effectiveness in detecting emotions in short texts is diminished because of its reliance solely on word frequency. The fifth sentence in Table V also shows the existence of ongoing trauma and fear toward men, making it more suitable to be categorized as fear. However, the use of the word "*dikecewain* (disappointed)" which can lead to the emotion of sadness. It causes misclassification by TF-IDF.

Sentence number seven exhibits errors in BERT's semantic features in Table V. The error by BERT is caused by a mixture of languages, such as Indonesian and Javanese, which cause errors in understanding the emotional context. Javanese is a regional language commonly used by people living on Java island in Indonesia. The Javanese phrase "*westerima kenyataan mbah* (just accept reality, grandma)" is unknown to the Indonesian-trained BERT model, potentially disrupting its semantic comprehension. However, models with statistical features like TF-IDF accurately predict emotions in mixed-language sentences despite disregarding sentence structure and syntax. TF-IDF disregards word relationships within sentences, focusing solely on word frequency in a document. For example, the word "*cemburuuuuu* (jealousyyyyy)" is recognized as a variation of "*cemburu* (jealousy)" and given equal weight. So, it aids in the proper classification of the emotion and makes it more resistant to the interference of mixed languages in a sentence.

Next, the researchers examine the effects of semantic and statistical aspects in brief and assess the ML algorithm's emotional detection competency, as displayed in Table VI. The result achieve the best performance based on the data from Table III. It includes semantic features from BERT on the Twitter dataset. The researchers analyz the Twitter dataset employing semantic features to evaluate various ML and DL techniques, including MLP, tree-based DT, probability-based NB, distance-based KNN, and LR.

DT is set up with a Gini criterion and max_depth of 7 [18]. The criterion function assesses the suitability of the split condition for guiding the input value toward the appropriate leaf node. The Gini index is selected for its ability to quantify a category's impurity in distinguishing an attribute [25]. The research employs Gaussian NB because of its compatibility with continuous vector input and Gaussian assumptions.

At the same time, KNN uses a parameter of eight neighbors based on the empirical results in previous research [26], indicating optimal performance in the 5–11 neighbors range. The number of nearest neighbors ($K$) or n_neighbors functions to predict categories in testing sample data [27]. Then, parameter used for LR is max_iter of 1000. It aims to set the maximum number of iterations. The optimization algorithm is carried out to find a convergent solution. A value of 1,000 indicates that the LR algorithm attempts up to 1,000 iterations to reach convergence. The parameters used for MLP are the same as those described previously.

In the evaluation results in Table VI, the DT algorithm displays an inferior F1-score of 0.56 for detecting emotions in short informal texts. DT algorithms struggle to generalize from informal text data, such as tweets and e-commerce product reviews, because

of their variability in language use, including slang and grammatical errors. The algorithms can mistake them for legitimate patterns. Moreover, the DT model excessively fits the training data. When compared to other studies, such as [12], which uses the statistical TF-IDF feature on the Twitter dataset [1] with two different algorithms, namely DT and KNN, it is only able to produce F1-scores of 0.51 and 0.49, respectively. Meanwhile, the research using the BERT semantic feature on the same dataset with a similar algorithm produces a higher F1-score, around 0.56 for DT and 0.64 for KNN. This comparison confirms that using BERT semantic features can improve classifier performance in emotion detection compared to the statistical TF-IDF feature due to BERT's ability to understand context and relationships between words through bidirectional learning [9].

Among other algorithms like DT, NB, KNN, and LR, the MLP algorithm achieves the best performance. MLP's advantage comes from its utilization of hidden layers with nonlinear activation functions, like ReLU, which enables the network to represent nonlinear relationships between features and labels. In informal texts, slang, abbreviations, and unstructured grammar can lead to variation and uncertainty. MLPs, due to their ability to model non-linear relationships, can capture complex patterns and interacting features that simple linear models cannot. MLP includes regulation techniques to minimize overfitting on noisy training data.

## IV. Conclusion

The research successfully analyzes the impact of statistical and semantic features on short texts, including tweets and Indonesian e-commerce product reviews. The research also considers ethics. During the data processing process, attention is given to privacy by deleting information that can reveal the identity of Twitter users or the names of reviewers' accounts on e-commerce platforms. Semantic feature extraction with BERT has been proven to provide better results than statistical feature extraction with TF-IDF in emotion detection in short texts. BERT can better understand sentence context because it produces high-dimensional vector representations which are rich in semantic information. It allows the model to capture the nuances and emotional context of short texts more effectively. Informal texts with emotionally similar keywords or nuances challenge TF-IDF, making them prone to misclassification due to their inability to grasp contextual relationships.

Several characteristics associated with short informal texts influence feature extraction outcomes, including (1) linguistic variants and slang, (2) language mixture, and (3) sentence context. First, informal texts often use nonstandard language, slang, and abbreviations, which are difficult for statistical methods such as TF-IDF to understand because these methods only look at word frequencies without considering the context. Second, language mixing can hinder BERT's contextual understanding if it is not adequately trained with diverse data. TF-IDF is more robust to interference because it determines word weight solely based on frequency, disregarding sentence structure or syntax. Third, BERT is superior in capturing the full sentence context, whereas TF-IDF can only view words independently. In addition, the MLP algorithm, a neural network-based DL algorithm, is more flexible and adaptive to variations and uncertainties in informal text because of its ability to model nonlinear relationships between features and labels. Regulation techniques in MLP also reduce overfitting, which is a common problem in algorithms like DT on varying training data.

The research is limited to comparing statistical and semantic features of short informal texts, not to the extent of comparing them with long formal texts such as news texts. Further research should be done to develop a more comprehensive dataset, which is applicable to lengthy formal texts, including mixed languages and regional dialects. Improving models and algorithms for text-based emotion detection in Indonesian can also be achieved using ensemble techniques that combine several models to increase prediction accuracy. In addition, language mixtures in sentences can be handled by developing multilingual models that can handle language mixtures more effectively or by data augmentation, which adds training data with mixed language sentences to train the model to understand language variations. Thus, the research results can provide practical value in business and product development for further research to prioritize handling product problems in e-commerce based on emotional intensity or developing product features based on emotional feedback. The results can also provide practical value in various social and organizational contexts, such as measuring public reactions to new policies or identifying issues that trigger strong emotions in society.



## Acknowledgement

The research was supported by a grant from Widya Husada University in 2023/2024. The authors are indebted to the Institute of Research and Community Outreach (LPPM) at Widya Husada University, which provided a grant to assist with this research.

Cite this article as: A. D. P. Ariyanto, F. K. Fikriah, and A. F. Setyawan, "Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences", CommIT Journal 19(1), 1–13, 2025.

## Author Contribution

Conceived and designed the analysis, collected the data, contributed data or analysis tools, and performed the analysis, A. D. P. A.; Wrote the paper, A. D. P. A., F. K. F., and A. F. S.

## Data Availability

The data that support the findings of the research are openly available in data.mendeley.com at https://doi.org/10.17632/574v66hf2v.1, reference number [4] and github.com at https://github.com/meisaputri21/Indonesian-Twitter-Emotion-Dataset, reference number [1].

## References

[1] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion classification on Indonesian Twitter dataset," in *2018 International Conference on Asian Language Processing (IALP)*. Bandung, Indonesia: IEEE, Nov. 15–17, 2018, pp. 90–95.

[2] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: A review of BERT-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[3] J. R. Andres, J. P. Soetandar, R. Sutoyo, and H. Riza, "Emotion recognition model using product review from Indonesia marketplace," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*. Banda Aceh, Indonesia: IEEE, Aug. 2–3, 2023, pp. 67–71.

[4] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, and S. M. Isa, "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks," *Data in Brief*, vol. 44, pp. 1–8, 2022.

[5] A. N. Sutranggono and E. M. Imah, "Tweets emotions analysis of community activities restriction as COVID-19 policy in Indonesia using support vector machine," *CommIT (Communication and Information Technology) Journal*, vol. 17, no. 1, pp. 13–25, 2023.

[6] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7819–7837, 2020.

[7] Y. Mao, L. Zhang, and Y. Li, "Finding product problems from online reviews based on BERT-CRF model," in *ICEB 2019 Proceedings*, Newcastle Upon Tyne, UK, 2019.

[8] I. K. Arsad, D. B. Setyohadi, and P. Mudjihartono, "E-commerce online review for detecting influencing factors users perception," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3156–3166, 2021.

[9] A. D. P. Ariyanto, D. Purwitasari, and C. Fatichah, "A systematic review on semantic role labeling for information extraction in low-resource data," *IEEE Access*, vol. 12, pp. 57 917–57 946, 2024.

[10] N. A. S. Winarsih and C. Supriyanto, "Evaluation of classification methods for Indonesian text emotion detection," in *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*. Semarang, Indonesia: IEEE, Aug. 5–6, 2016, pp. 130–133.

[11] A. D. P. Ariyanto, C. Fatichah, and D. Purwitasari, "Semantic role labeling for information extraction on Indonesian texts: A literature review," in *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2023, pp. 119–124.

[12] K. S. Nugroho, F. A. Bachtiar, and W. F. Mahmudy, "Detecting emotion in Indonesian Tweets: A term-weighting scheme study," *Journal of Information Systems Engineering & Business Intelligence*, vol. 8, no. 1, pp. 61–70, 2022.

[13] A. Nurkasanah and M. Hayaty, "Feature extraction using lexicon on the emotion recognition dataset of Indonesian text," *Ultimatics: Jurnal Teknik Informatika*, vol. 14, no. 1, pp. 20–27, 2022.

[14] H. T. Duong and T. A. Nguyen-Thi, "A review: Preprocessing techniques and data augmentation for sentiment analysis," *Computational Social Networks*, vol. 8, pp. 1–16, 2021.

[15] A. D. P. Ariyanto and A. Z. Arifin, "Analisis metode representasi teks untuk deteksi interelasi Kitab Hadis: Systematic literature review," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 5, pp. 992–1000, 2021.

[16] A. D. A. P., A. A. Z., R. S. W., and R. I., "Metode pembobotan kata berbasis cluster untuk perangkingan dokumen berbahasa Arab (Cluster-based word weighting method for ranking Arabic documents)," *Techno.COM*, vol. 20, no. 2, pp. 259–267, 2021.

[17] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," 2020. [Online]. Available: https://arxiv.org/abs/2003.07278

[18] A. D. P. Ariyanto, "Deteksi interelasi antar Kitab Hadis menggunakan word embedding dan ensemble learning," Master's thesis, Institut Teknologi

Sepuluh Nopember, 2022.

[19] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *Journal of Big Data*, vol. 8, pp. 1–16, 2021.

[20] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, "Colloquial Indonesian lexicon," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 226–229.

[21] J. Santoso, A. D. B. Soetiono, E. Setyati, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Self-training naive bayes berbasis Word2Vec untuk kategorisasi berita bahasa Indonesia," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 7, no. 2, pp. 158–166, 2018.

[22] F. Z. El-Alami, S. O. El Alaoui, and N. E. Nahnahi, "Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 8422–8428, 2022.

[23] B. M. Hsu, "Comparison of supervised classification models on textual data," *Mathematics*, vol. 8, no. 5, pp. 1–16, 2020.

[24] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19 083–19 095, 2022.

[25] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020.

[26] N. A. P. Rostam and N. H. A. H. Malim, "Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 658–667, 2021.

[27] H. A. Abu Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. Eyal Salman, and V. S. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: A review," *Big Data*, vol. 7, pp. 221–248, 2019.