**Indonesian Journal of Electrical Engineering and Informatics (IJEEI)**
Vol. 10, No. 1, March 2022, pp. 219~231
ISSN: 2089-3272, DOI: 10.52549/ijeei.v10i1.3565


# Sentiment Analysis in Karonese Tweet using Machine Learning

**Ichwanul Muslim Karo Karo[1], Mohd Farhan Md Fudzee[2], Shahreen Kasim[3] & Azizul Azhar Ramli[4]**
[1,2,3,4]Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn, Malaysia

**Article Info**

***Article history:***

Received Dec 14, 2021
Revised Feb 24, 2022
Accepted Mar 8, 2022

***Keyword:***

Karonese
Sentiment analysis
Logistic regression
Naïve Bayes
K-nearest neighbor
Support Vector Machine

**ABSTRACT**

Recently, many social media users expressed their conditions, ideas, emotions using local languages on social media, for example via tweets or status. Due to the large number of texts, sentiment analysis is used to identify opinions, ideas, or thoughts from social media. Sentiment analysis research has also been widely applied to local languages. Karonese is one of the largest local languages in North Sumatera, Indonesia. Karo society actively use the language in expression on twitter. This study proposes two things: Karonese tweet dataset for classification and analysis of sentiment on Karonese. Several machine learning algorithms are implemented in this research, that is Logistic regression, Naive bayes, K-nearest neighbor, and Support Vector Machine (SVM). Karonese tweets is obtained from timeline twitter based on several keywords and hashtags. Transcribers from ethnic figures helped annotating the Karo tweets into three classes: positive, negative, and neutral. To get the best model, several scenarios were run based on various compositions of training data and test data. The SVM algorithm has highest accuracy, precision, recall, and F-1 scores than others. As the research is a preliminary research of sentiment analysis on Karonese language, there are many feature works to improvement.



***Corresponding Author:***

Ichwanul Muslim Karo Karo
Faculty of Computer Sciences and Information Technology,
Universiti Tun Hussein Onn, Malaysia,
Persiaran Tun Dr. Ismail Street Kluang Parit Raja, 86400 Batu Pahat, Johor, Malaysia
Email: farhan@uthm.edu.my

## 1. INTRODUCTION

The industrial revolution 4.0 and the increasing number of social media users have fundamentally changed the way people express opinions and share information [1]. The rapid growth and popularity of social media networks is linear with the number of diverse social media users and an abundance of user-generated content. The diversity of social media users affects the language used on social media, for example one of the social media active user habits is using their local language or national language in expressing a status, tell stories or vent emotion than using English as international language. It is because they feel more comfortable and easier to understand for the user's circle of friends. Thus, the abundance of text data from social media is an opportunity to identify and determine opinions and perspectives from content so as to increase the value of various fields, such as economics, social, politics, business and others in a particular language user.

Twitter is the top five most popular social media platforms [2, 3] where in a minute there are many tweets are shared and read many people. Tweets could contain valuable data that represented some people situation. Recently, Twitter has become a place to share opinions, ideas, emotions, or perspectives on an event and so on. Twitter accommodates tweets in various languages on alphabetic. In other words, expression on Twitter is not only in English, but in local or other national languages. As a result, there are tons of tweets in multiple languages being shared every minute.

entiment analysis is a research area to overview people's attitudes, emotions, and opinions from text, by involves a combination of text mining and natural language processing [6]. Nowadays, sentiment analysis does

not only analyzes English text, but it has been implemented in many local languages, such as Javanese [7], Sundanese [8], Minangkabau [9], Tamil [2], Malayalam [10], and Afaan Oromoo [1] or nationality languages such as Arab [6], Indonesia [4] , Chinese [11], Turkish [12], Azerbaijani [5]. Sentiment analysis is commonly used to identify the positive and negative or neutral feelings within text of a language on social media. Application of sentiment analysis to transform the abundance of information in the form of text and active users on social media into actionable insights.

Karo is one of the largest tribes in North Sumatra in Indonesia [13]. The Karo language or the Karo Batak language belongs to the Austronesian language family used by the Karo society [14]. The Karo people use Karonese as a communication tool to convey ideas, thoughts, intentions, and goals. They also use Karonese to communicate or express on social media, such as posting a status, tell stories or vent emotion. There is a lot of Karonese text data on social media. This is an opportunity for research in the sentiment analysis field. Besides, this is also a challenge, considering that until now there has been no research on sentiment analysis for the Karonese, so the availability of data and previous work is limitation.

There are two fundamental approaches to solve sentiment analysis on local or non-English language, these are dictionary-based, machine learning and hybrid [15]. Dictionary-based sentiment analysis uses a dictionary of words that are labeled by scoring. Machine learning approach generally uses a text classification algorithm to identify the pattern on text. Hybrid method is combining both approaches. However, sentiment analysis on local or non-English languages has a bigger challenge than English [9, 16]. There are several reasons for this, firstly, supply English text data is more wealthy than other text data [2, 16], second, data preparation of English text is clearly [6, 11], such as library stemming, lemmatization, and Stop-word removal. Third, limited data training (text and label) to analyze sentiment using machine learning approach and the effort and cost of dictionary approach is very large for non-english text [16].

Machine learning approach is an option that is often used to analyze non-English sentiments (local or other nationality language). Machine learning is widely used for multilingual sentiment analysis. Study [12] analyzed sentiment in Turkish using four machine learning methods, Naïve Bayesian (NB), Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN). The best accuracy from his research was obtained by using K-NN. Another research was conducted by [6], they conducted Arabic Language Sentiment Analysis on Health Services using many machine learning methods, the SVM algorithm was superior to other algorithms in their experiments. Another study analyzed Azerbaijani tweets using Logistic Regression, Naive Bayes and SVM [5]. In the research, the combination of TF-IDF and SVM outperforms other methods and combinations.

Machine learning approach has also been applied to analyze sentiment in local languages. Study by [8] analyzed sentiment Sundanese (local language from Indonesia) using combination Naïve bayes and Chi Squared Statistics. Combination of Naïve bayes algorithm with tf-idf also has been implemented to analyze sentiment on Malayalam (South India language) tweets [17] . Further research is still from the south Asia, a study by [18] compared several machine learning algorithms to analyze sentiment Tamil text (etnic from south Asia). SVM algorithm outperform to other algorithms in the case. The next sentiment analysis for local language comes from the country of Ethiopia, Afaan Oromo language was analyzed using Naïve bayes with tf-df [19]. Based on several of these studies, the machine learning approach is more familiar to be applied for preliminary research on sentiment analysis on local or non-English languages.

This research proposes two things: Karonese tweets dataset for sentiment analysis and implemented using machine learning. The proposed work explains sentiment analysis of Karonese tweets, which have been classified into positive, negative, or neutral using different machine learning algorithms like NB, SVM, K-NN and Logistic regression (LR). The rest of the paper is organized as, section 2 describes about the research methods and briefs dataset, preprocessing and machine learning model used for performing experiment. The experiment and results are discussed in section 3. Conclusion and feature work present in section 4.

## 2. RESEARCH METHOD

Figure 1 shows the research flowchart of this paper. Research consists of several important stages. The process begins with retrieving the Karonese tweets. The next stage is text pre-processing, this process will adjust to recent preprocessing techniques. To complete the training data for sentiment analysis, each tweet will be labeled with a class positive, negative, or neutral. The annoted process is assisted by ethical group figures. The next process is to classify Karonese tweets using a machine learning algorithm. The last process of this research is to evaluate the performance of each algorithm.

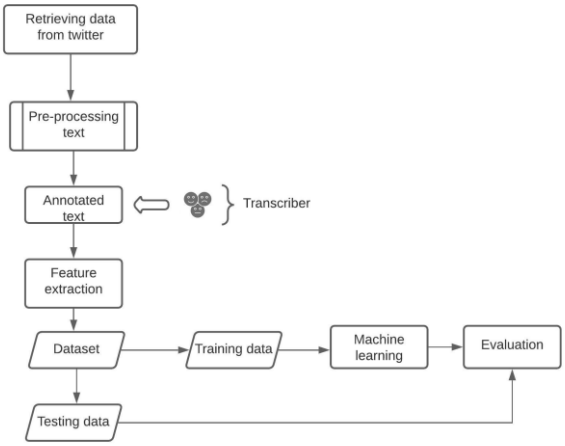


Figure 1. Researh flowchart

### 2.1. Dataset

Karonese text was crawled from timeline Twitter between 01/01/2021 to 31/10/2021 using the API and python programing. Karonese tweets contain ideas, opinions, or reviews. Karonese text crawling process was run twice. The first crawling process used several keywords related to the Karonese, such as "deleng sinabung, Sinabung mountain", "mejuah-juah, greeting welcome", "Gundaling", and so on. However, insufficient number of tweets obtained using keywords. The second crawling process is based on several hashtags such as #kalakkaro, # #antonyginting, and #lyodra. There are at least three features that have been successfully crawled, that is tweets, number of likes and retweets. This research only used Karonese's tweets and comments. Meanwhile, other features and non-Karo language comments are not processed. The number of tweets that were successfully obtained was 780, the data snippet is shown in Table 1. Lastly, tweets stored on .csv format.

Table 1. Example of dataset

| No | Tweet | like | retweet |
|---|---|---|---|
| 1 | Ija pa pe kam ringan gundari, kutotoken gelah kam sehat-sehat | 2 | 4 |
| 2 | Kai Lagu Karo popular gundari nake? | 0 | 0 |
| 3 | rubati ras kam perban ndekahsa kam ciloak i pajak singa | 0 | 0 |
| 4 | Deleng Sinabung Meletus mulihi | 3 | 10 |
| 5 | Abu deleng she ku gundaling | 1 | 2 |

### 2.2. Pre-processing

Text preprocessing plays an important role in analyzing tweets [20]. Text preprocessing aims to convert unstructured data into structured data using natural language processing (NLP). The transformation process adapts to the needs of the mining process (sentiment analysis, summarization, document clustering, etc.). Concisely, preprocessing is converting text into index terms. The goal is to generate a set of index terms that can represent a document. There are many processes on recent NLP [21], such as tokenization, Stop-word removal, symbolic removal, feature selection, negation N-gram, POS tagging, noise removal, stemming, feature extraction, machine translation, lemmatization, and capitalization. However, However, not all NLP processes apply to certain languages. NLP requires a special stemmer algorithm for certain languages [22, 23], distinct stop-word removal library for each language, distinct lemmatization algorithm for a specific language [24] and other distincts.

As preliminary research on Karonese sentiment analysis, of course there are many limitations to text processing techniques. Among the text preprocessing techniques that have not been applied are Stemming, Stop-word removal and lemmatization. The text preprocessing techniques applied in this work, casefolding, symbolic removal, tokenization, and feature extraction. Table 2 is an example of Karonese text preprocessing input and output. The next stage of text processing is feature extraction (explain in subsection 2.4).

Table 2. I/O text preprocessing

| Input text | Text processing techniques | Output |
|---|---|---|
| Kai Lagu Karo popular gundari nake? | Case folding | kai lagu Karo popular gundari nake? |
| kai lagu Karo popular gundari nake? | Symbolic removal | kai lagu karo popular gundari nake |
| kai lagu Karo popular gundari nake | Tokenization | "kai", "lagu", "karo", "popular", "gundari", " nake" |

### 2.3. Labelling Dataset

Due to limited resources, several researchers complement their own non-English language data set by manually labeling. Study by [25] build labeled Afaan Oromo language dataset by manually annotating 1810 sentences taken from Ethiopian governmental broad casting cooperation, Oromia broad casting network (OBN) twitter page. In other studies, a paper by [26] also build labeled dataset by manually annotating Bambara sentences. The sentences obtained from Malian Facebook users. The other datasets are Punjabi news articles [27], Indonesian review from KitaReview website [28], comments/reviews from Vietnamese commercial web pages and was annotated by three human annotators, Jordanian dialect tweets [3] and and other languages.

To complete the Karonese dataset, sentences will also be manually labeled. The tweet labeling process aims to identify tweet into positive, negative, or neutral classes. This process is assisted by four expert transcribers from ethinc figure. Each transcriber labels the same tweets, so that a tweet gets four labels from the transcript, a positive, negative, or neutral label. To determine the final label of a sentence, the plurality method is used [29]. Plurality method is only chosen first place votes (most votes win = plurality candidate). More formally, let there be $n$ transcribers $T = \{T_1, T_2, T_3 \ldots T_n\}$, $k$ sentences $S = \{S_1, S_2, S_3 \ldots S_k\}$, and $m$ C $=\{C_1, C_2, C_3 \ldots C_n\}$ candidate class. Each transcriber $T_i$ label $S_i$ into $C$ according to his or her preferences, creating a preference class $\Omega$, where $\Omega$ is the set of all possible preference class. $S_i$ is class $C_1$, if $|C_1| \geq |C_2| \geq |C_3| \ldots |C_n|$ from $\Omega_i$. Finaly, we have 209, 311 and 260 for positive, negative, and neutral class (Figure 2.).

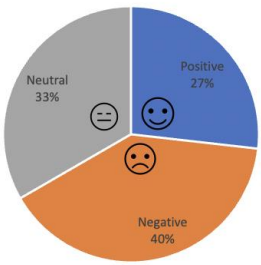


Figure 2. Composition of class

### 2.4. Feature Extraction

The high dimension of the feature space is the major issue in text classification. Many of these features aren't appropriate or effective for text classification. The classification accuracy can be severely harmed by certain noise features. Furthermore, many features can stutter the classification process or even render some classifiers ineffective. Consequently, text classification applies feature selection to reduce the size of the feature field and improve the efficiency and accuracy of classifiers. To examine preprocessed data, it should be translated to features. In this investigation, we used TF-IDF to extract the feature.

The TF-IDF (Term frequency-inverse document frequency) technique is a text processing technique that is used to analyze the most essential terms in a document [30]. The TF-IDF score is formed by two methods: TF (term frequency) and Idf (invers document frequency). TF-IDF can be calculated using equation (1).

$$w(i,j) = TF_{i,j}.log\frac{N}{df_i} \quad (1)$$

Where $TF_{i,j}$ is number occurance term $i$ on $j$, $df_i$ represent number of documents containing term $i$ and $N$ is total number of documents.

### 2.5. Machine Learning

The next process is to build a sentiment analysis model using machine learning methods. There are four different machine learning (ML) methods, namely Logistic regression (LR), Naïve Bayesian (NB), and Support vector Machine (SVM) used in this research. Each algorithm will be tested independently with various scenarios.

a. Logistic Regression (LR)

Logistic regression is another powerful supervised ML algorithm used for classification problems especially target is categorical [31]. The idea of logistic regression is a linear regression for classification problems. A logistic function (equation 2) is used to predict a binary output variable in logistic regression [5]. The main distinction between linear and logistic regression is that the range of logistic regression is limited to 0 and 1. Furthermore, logistic regression does not require a linear relationship between input and output variables, unlike linear regression.

LR is a parametric form for the distribution P(Y|X) where Y is a discrete value and $X = \{x_1, x_2, x_3 \ldots, x_n\}$ is a vector containing discrete [32]. Let Y = {-1,0,1}, the parametric model of LR can be written as equation (2). The predictions will be categorized as class 0 if the probability is larger than 0.5. Class 1 will be assigned if this does not result.

$$Logistic\ function\ (Y) = \frac{1}{1 + e^{-x}} \tag{2}$$

b. Naïve Bayes (NB)

The Naive Bayes algorithm is an algorithm for classifying data into the most appropriate category by determining the highest probability value [8]. When applied to big databases, the Nave Bayes classifier has great accuracy and speed. The Nave Bayes method is based on Thomas Bayes's theorem, which was established in the 18th century. The formula shows on equation (3).

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{3}$$

The probability of hypothesis X based on the condition of H is known as likelihoo (*P (X | H)*. Meanwhile, P *P (H)* and *P (X)* represent the likelihood of texting data with an unknown class and the chance of texting data with a given class, respectively. The Naive Bayes theorem adjusts to numerous conditions in text classification applications. Let, $j$ is text categories ($Cj$), each $X$ contains word ($w_i$) and assume that each word in category is independent, so the Naïve Bayes calculation can be simplified further as equation (4)

$$P(X|C_j) = \prod_{i=1}^{n} P(w_i|C_j) \tag{4}$$

Meanwhile, to determine the class of a tweet, it is obtained from the maximum result value $P(w_i|C_j)$ (probaliblitas word *i* on class *j*) and probalibilitas class *j* ($P(C_j)$) in equation (5). Then Naïve bayes algorithm shown on Figure

$$C_{tweet\ j} = argmax\ P(C_j).\prod_{i=1}^{n} P(w_i|C_j) \tag{5}$$

**Naïve Bayes Algorithm**

1. Prepare a dataset
2. Count the number of classes in the training
3. Count the same number of cases with the same class
4. Multiply all results according to the testing data that the class will look for
5. Compare outcomes per class with the highest value being assigned to the most recent class

Figure 3. Naive bayes algorithm

c. Support Vector Machine (SVM)

**SVM Algorithm**

**Input:**

$N_{in}$(The number of input vector)
$N_{sv}$(The number of support vector)
$N_{ft}$(The number of features in support vector)
$SV[N_{sv}]$ (Support vector array)
$IN[N_{in}]$ (Input vector array)
$b*$ (bias)

**Output:**

$F$ (decision function output)

```
for i ← 1 to N_in by 1 do
    F = 0
    for j ← 1 to N_sv by 1 do
        dist = 0
        for k ← 1 to N_ft by 1 do
            dist += (SV[j].feature[k] − IN[i].feature[k])²
        end
        k = exp (−γ × dist)
        F+= SV[j].a* × k
    end
    F = F + b*)
end
```

Figure 4. SVM algorithm

Support Vector Machines (SVM) is a machine learning algorithm that is capable of dealing with multiple variables and classes [33]. SVM are learning methods that use the generalization theory to efficiently train linear learning machines in kernel-induced feature spaces [34]. Due to the Karush–Kuhn–Tucker criteria, SVM generate a sparse dual representation of the built hypothesis, resulting in efficient learning algorithms that can be solved using an optimization method. Furthermore, the issue is convex, and the solution converges to a global optimum value. These characteristics distinguish SVM from other pattern recognition approaches like neural networks and the Decision Tree Algorithm. The SV learning machine's goal is to determine f(x,α), which α corresponds to the weights and biases in order to map the underlying link between input drivers and output responses. The SVM algorithm trains machines by minimizing an upper bound on the generalization error. Traditional learning techniques such as neural networks minimize training error on training data, whereas the SVM algorithm trains machines by minimizing an upper bound on the generalization error. As a result, rather than minimizing the empirical risk, SVM learning focuses on decreasing the structural risk. In other words, SVMs try to reduce the chance of misclassifying test data that isn't visible to the model and is taken at random from a known but unknown probability distribution. Figure 4 shows the fundamental SVM method.

d. K-nearest neighbor

K-Nearest Neighbor (K-NN) algorithm is a populer supervised algorithm that uses distance function to classify data [35], the algorithm shown on Figure 5. Similarity methods play important rule in this algorithm. The basic idea behind the technique is to calculate object similarity and arrange them into the highest similarity class. Finding the $K$ group of items in the training data that are closest (similar) to the object in new data or data testing is the ultimate state of K-NN. The Euclidean distance is the most famous distance function used in data analysis.

**K-NN Algorithm**

**Input:**

Training sample $D$, test sample $d$, $K$

**Output:**

Class label of test sample

1. Compute the distance between $d$ and every sample in $D$
2. Choose the $K$ samples in $D$ that are nearest to $d$; denote the set by $P$ ($\in D$)
3. Assign $d$ the class that is the most frequent class (or the majority class)

Figure 5. K-NN algorithm

### 2.6. Evaluation

Lastly, Output of text classification is a classification model. The model will be tested against train data and data testing and evaluated using several parameters. As evaluation metrics: accuracy, precision, recall, and F1-scores are calculated from the confusion matrix (Table 3).

Table 3. Confusion matrix

| | | Predict | |
|---|---|---|---|
| | | Positive | Negative |
| Actual | Positive | True positive (TP) | False negative (FN) |
| | Negative | False positive (FP) | True Negative (TN) |

The ratio of successfully predicted positive observations to total predicted positive observations is known as precision (P) [35]. The formula shown in the equation (6).

$$P = \frac{TP}{TP + FP} . 100\% \quad (6)$$

The ratio of precisely predicted positive observations to the total number of relevant samples is known as recall (R) [35]. The formula to calculated shown in the equation (7).

$$Recall(R) = \frac{TP}{TP + FN} . 100\% \quad (7)$$

F1 score, which is shown in equation (8), is the weighted average of Precision and Recall and a method used to measure the performance of the model.

$$F1\ score = \frac{2PR}{P + R} . 100\% \quad (8)$$

The percentage of correctly classified instances is called accuracy, which is shown in equation (9). The most intuitive performance metric is accuracy, which is just the ratio of properly predicted observations to all observations.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} . 100\% \quad (9)$$

## 3. RESULTS AND DISCUSSION

In this section, the results of research were explained and at the same time is given the comprehensive discussion. Each machine learning algorithm was ran using various compositions of trining data and data as experimental scenarios (Table 4.). The composition of the data is obtained based on recent sentiment analysis research [8, 11, 35] . In addition, this section will also compare the performances between algorithms.

Table 4. Experimental scenarios

| Scenario | Data composition (training : testing) |
|---|---|
| I | 80:20 |
| II | 60:40 |
| III | 50:50 |
| IV | 40:60 |

### 3.1. Karonese tweet sentiment analysis using LR algorithm with TF-IDF

The first experiment uses the Logistic regression approach with TF-IDF to analyze sentiment on Karonese. There were four situations run tested: I, II, III, and IV (Table 4). Until this journal is published, no one has completed a sentiment analysis on Karonese using Logistic Regression with TF IDF using any data. Table 5 displays the results of each test. In comparison to other models, the scenario III model had the highest accuracy, precision, recall, and f-1 scores, as shown in the table. In general, the model from scenario III performs better than 50% of the time. Obviously, this conclusion is unsatisfactory, but as a preliminary sentiment analysis study on Karonese, it can be used to propose a new idea or improve an existing one.

Table 5. Performance of logistic regression algorithm with TF-IDF

| Scenario experiment | Accuracy | Precision | Recall | F-1 score |
|---|---|---|---|---|
| I | 44.8 | 44.7 | 45.2 | 44.9 |
| II | 49.5 | 50.2 | 51.4 | 50.8 |
| III | 51.9 | 54.6 | 52.2 | 53.4 |
| IV | 41 | 41.8 | 38.7 | 40.2 |

The drawbacks of Logistic Regression are demonstrated in this study. With high-dimensional datasets, logistic regression techniques have significant issues [5, 31]. There are 2104 features in this investigation dataset. Due to the obvious large dimension of the Karonese dataset, we suspect that the logistic regression model's performance is not optimal, if not poor. This fortifies the assertion [31] that Logistic Regression should not be employed if the number of observations is smaller than the number of features. The difficulty in interpreting logistic regression models is another flaw. It's because the weights' meaning is multiplicative rather than additive. The usage of static weights is the second cause of the non-optimal logistic regression technique in this scenario.

### 3.2. Karonese tweet sentiment analysis using NB algorithm with TF-IDF

The second experiment is a sentiment analysis on Karonese using the Naïve Bayes algorithm with TF IDF. In this experiment, four test scenarios were ran, the same as the previous experiment. This algorithm has also never been implemented to analyze sentiment on Karonese. In other words, for the first time the Naïve Bayes algorithm is used to analyze sentiment on Karonese. The results of each test can be seen in Table 6. Based on the table, the best performance was obtained from the scenario I models, the best precision, recall and F-1 score were obtained from scenario I. Unfornutely, overall, of the performances of the Naïve Bayes model is below 50 percent in analyzing sentiment on Karonese.

Table 6, Performance of Naive bayes algorithm

| Scenario experiment | Accuracy | Precision | Recall | F-1 score |
|---|---|---|---|---|
| I | 51 | 50.1 | 50.6 | 50.3 |
| II | 33.5 | 27.4 | 39.5 | 32.4 |
| III | 45 | 45.2 | 43.1 | 44.1 |
| IV | 20.4 | 12.8 | 32.7 | 18.4 |

Researchers analyze those words in Karo are interrelated features. For example, the word "jong, corn" does not describe any sentiment, but if it is paired with the word "macik, rotten", it will have a negative sentiment. We argue that these conditions make the Naïve Bayes algorithm weak in predicting dependent features and obtaining bad quality results. Our statement is also corroborated by previous research [8, 31, 4], which reveals that Naïve bayes is a bad choice algorithm to classify dependent features.

In addition, The Nave Bayes algorithm also has big problems with sparse datasets [31]. Karonese dataset consists of 780 rows (number of tweets) and 2104 features (uniques term), so there are many features on tweets with a value of 0 (Table 7.). This study also confirms that Naïve Bayes has problems with data containing sparse vectors. Based on the sentiment results and characteristics of the Naïve Bayes algorithm, we argue that if the Naïve Bayes algorithm is forced to analyze the sparse dataset, the results of the performances are not optimal.

Table 7. Illustration of sparse dataset on Karonese

| Tweet_id | aku | enggo | man | … | … | … | jong | macik |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

### 3.3. Karonese tweet sentiment analysis using SVM algorithm with TF-IDF

The next experiment is sentiment analysis on Karonese using the SVM with TF-IDF algorithm. The SVM algorithm has also never been used to analyze sentiment on Karonese. This experiment also uses four scenarios. The parameters used are linear kernel functions and degree = 3. The experimental results can be seen in Table 8. The best accuracy, recall and F-1 scores were obtained from the scenario I model, while the best precision was obtained from the scenario II. The performance of the scenario I model is mor e stable than the other scenarios. This can be seen from the distribution of the value of performance parameters which is evenly distributed with performances above 50 percent.

Table 8. Performance of SVM algorithm

| Scenario experiment | Accuracy | Precision | Recall | F-1 score |
|---|---|---|---|---|
| I | 57 | 57.3 | 56.4 | 56.8 |
| II | 52.7 | 59.8 | 44.7 | 51.2 |
| III | 37 | 29.1 | 40.8 | 34 |
| IV | 20 | 12.8 | 30.1 | 18 |

### 3.4. Karonese tweet sentiment analysis using K-NN algorithm with TF-IDF

The last experiment is a sentiment analysis on Karonese using the K-NN with TF-IDF algorithm. The K-NN algorithm has also never been used to analyze sentiment on Karonese. This experiment also uses four scenarios, with number of $K = 3$ [35]. The experimental results can be seen in Table 9. The best performance was obtained from the scenario I model. The performance of the scenario I model is more stable than the other scenarios. This can be seen from the distribution of the value of performance parameters which is evenly distributed with performances above 50 percent.

Table 9. Performance of K-NN algorithm

| Scenario experiment | Accuracy | Precision | Recall | F-1 score |
|---|---|---|---|---|
| I | 52 | 53.3 | 50.1 | 51.7 |
| II | 40 | 41.5 | 39.8 | 40.6 |
| III | 50.1 | 52.9 | 47 | 45.8 |
| IV | 37 | 34 | 38 | 36 |

### 3.5. Comparison analysis of algorithms

This section presents a comparison of the four algorithms and discuss them. Each algorithm uses a different approach in obtaining the classification model. Logistic regression uses a predictive approach, Naïve Bayes algorithm uses a probability approach while SVM and K-NN uses a similarity approach between objects. The performance comparison of the four algorithms can be seen in Figure 6. In general, the performance between algorithms is not very significant, the algorithm performance is distributed around 50 percent. SVM's accuracy, precision and F-1 score algorithm are superior to other algorithms. While the best recall is obtained by the SVM, KNN and Logistic regression algorithms. The Naïve Bayes algorithm is a bad choice for analyzing sentiment on Karonese. Furthermore, we see that the performance gap of SVM versus KNN is the same as the performance gap of the SVM algorithm versus logistic regression.

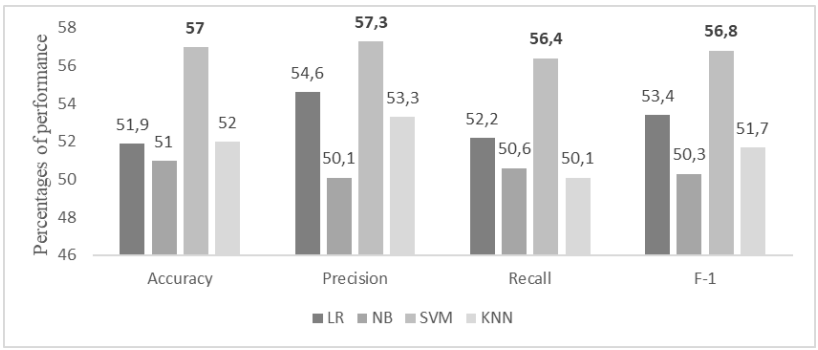


Figure 6. Comparison of machine learning algorithms

The results of the sentiment analysis on Karonese are not quite satisfactory. As preliminary research on sentiment analysis for Karonese, we see that there are two important points that are not optimal. First, recent preprocessing techniques cannot be fully applied on Karonese text. There are many the existing techniques does not work well, such as steaming, lemmatization, stopword removal and soon. This is because there have

not been many studies that proposed preprocessing techniques on Karonese to solve it. The next process is that traditional machine learning is not capable of analyzing Karo language sentiment, so tweets are not well separated. This is reinforced by the results of the classification, there is no single algorithm with performances above 70 percent.

### 3.6. Comparison analysis with other studies

This sub section presents recent research on Karonese. The aim is to show the gap of existing Karonese research with this research. So far, Karonese's research still focuses on text preprocessing (Table 10), such as stemming, identifying morphology and phonology on Karonese and translating Karonese into English. Translation of Karonese text to English could be an alternative to provide good resources for good Karonese sentiment analysis. This research is a baseline to improve Karonese sentiment analysis results.

Table 10. Recent Karonese research

| Related study object | Topic research | Limitations for Karonese sentiment analysis |
|---|---|---|
| [36] | Karonese language text stemming algorithm based on grammar | The research has provided a text preprocessing technique based on a document. However, the research does not present a dataset for KSA until testing on KSA |
| [37] | The translation of "*Ngangkat Tulan-Tulan"* texts in Karonese society into English | The process of translating Karonese text to English aims to provide KSA datasets. However, the process of karo language is not accommodated by machine translation and cultural vocabulary does not have the right equivalent. |
| [38] | Derivational morphology of Karonese ecolexicon | Karonese is a unique morphology language. This study identifies Karonese morphology that is useful in Karonese text, but has not normalized every word |
| [39] | Phonological dialect differences of karonese language in Medan, north Sumatra | Karonese has unique phonology. This work identifies Karonese phonology. The knowladge is useful for preprocessing Karonese text, but it doesn't normalize every word |
| [40] | Translating Textual Theme in Maba Belo Selambar Dialogue of Karonese Society into English | The results of translated sentences can be used as a sentiment analysis dataset, but the class for the sentence is not yet available. |

Sentiment analysis has been widely applied on many non-English languages (local or other national languages). A different language is a novelty in sentiment analysis on non-English research [16] [21]. Karo language as a language object has never been used in sentiment analysis (Table 11). As a preliminary Karonese sentiment analysis using machine learning, Table 10 presents the basic machine learning algorithms for several non-English sentiment analysis. Based on the table, the SVM algorithm is often the choice as a sentiment analyzer for non-English SA. In addition, the SVM algorithm also provides the best accuracy compared to other algorithms.

Table 11. Comparison analysis of machine learning algorithm on preliminary non-English sentiment analysis

| Related work | Algorithms | Accuracy |
|---|---|---|
| Arabic Jordanian SA [3] | NB | 0.5 |
| | SVM | **0.66** |
| Azerbaijani SA [5] | NB | 0.90 |
| | SVM | **0.93** |
| | LR | **0.93** |
| Afaan oromo SA [19] | SVM | **0.90** |
| | Random Forest | 0.89 |
| Bambara SA [41] | SVM | **0.71** |
| | LR | 0.68 |
| **This work** | SVM | **0.57** |
| | LR | 0.519 |
| | NB | 0.51 |
| | KNN | 0.52 |

## 4. CONCLUSION

A sentiment analysis study on Karonese has been started in this paper. The basic techniques of sentiment analysis have been applied, including data collection, preprocessing, and classifying tweets in the

Karo language. On the preliminary issue of the TF-IDF model, machine learning methods such as logistic regression, Nave Bayes, SVM, and KNN are used to detect text sentiment. In comparison to other algorithms, the SVM method has been proposed to produce better results in experimental experiments, with an accuracy of 57%, precision of 57.3 percent, recall of 56.4 percent, and F-1 score of 56.8%. Furthermore, the majority of the best models were derived from the data composition by comparing the training and testing data in an 80:20 ratio.. Based on unsatisfactory results, would suggest that preprocessing tweets and methods may need further improvement or modification.

## ACKNOWLEDGEMENTS

This work is supported by the ministry of higher education (MOHE) under the fundamental research grant scheme (FRGS) reference code FRGS/1/2018/ICT04/UTHM/02/3.

## REFERENCES

[1] M. O. Rase , "Sentiment Analysis of Afaan Oromoo Facebook Media Using Deep Learning Approach," *New Media and Mass Communication,* vol. 90, pp. 7-22, 2020.

[2] S. Anbukkarasi and S. Varadhaganapathy, "Analyzing Sentiment in Tamil Tweets using Deep Neural Network," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE*, 2020.

[3] J. O. Atoum and M. Nouman, "Sentiment analysis of Arabic jordanian dialect tweets," *Int. J. Adv. Comput. Sci. Appl,* vol. 10, no. 2, pp. 256-262, 2019.

[4] V. A. Fitri, R. Andreswari and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Computer Science,* vol. 161, pp. 765-772, 2019.

[5] H. Hasanli and S. Rustamov, "Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM," in *2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)* , 2019.

[6] A. M. Alayba, V. Palade, M. England and R. Iqbal, "Arabic Language Sentiment Analysis on Health Services," in *2017 IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR)*, 2017.

[7] C. Tho, Y. Heryadi, L. Lukas and A. Wibowo, "Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach," *Journal of Physics: Conference Series,* vol. 1869, no. 1, 2021.

[8] Y. Cahyono and S. Saprudin, "Analisis Sentiment Tweets Berbahasa Sunda Menggunakan Naive Bayes Classifier dengan Seleksi Feature Chi Squared Statistic," *Jurnal Informatika Universitas Pamulang,* vol. 4, no. 3, pp. 87-94., 2019.

[9] F. Koto and I. Koto, "Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation," in *The 34th Pacific Asia Conference on Language, Information and Computation*, 2020.

[10] S. S. Kumar, M. A. Kumar and K. P. Soman, "Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets," in *International Conference on Mining Intelligence and Knowledge Exploration. Springer*, 320-334.

[11] X. Lin and C. Han, "Chinese Text Sentiment Analysis Based on Improved Convolutional Neural Networks," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018.

[12] M. Rumelli and D. Akkuş, "Sentiment Analysis in Turkish Text with Machine Learning Algorithms," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, 2019.

[13] B. Tarigan, R. Sofya and R. N. Rosa, "Derivational Morphology of Karonese Ecolexicon," in *Proceedings of the Seventh International Conference on Languages and Arts (ICLA 2018)*, 2019.

[14] G. Woollams , A GRAMMAR OF KARO BATAK, SUMATRA, Australia: Australian National University Canberra , 1996.

[15] B. Verma and R. S. Thakur, "Sentiment analysis using lexicon and machine learning-based approaches: A survey," in *Proceedings of international conference on recent advancement on computer and communication. Springer„* Singapure, 2018.

[16] F. Djatmiko, R. Ferdiana and M. Faris, "A Review of Sentiment Analysis for Non-English Language," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, 2019.

[17] S. Soumya and K. V. Pramod, "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express,* vol. 6, no. 4, pp. 300-305., 300-305..

[18] S. Se, R. Vinayakumar, M. A. Kumar and K. P. Soman, "Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms," *Indian Journal of Science and Technology,* vol. 9, no. 45, pp. 1-5.

[19] M. Oljira, "Sentiment Analysis of Afaan Oromo using Machine learning Approach," *International Journal of Research Studies in Science, Engineering and Technology,* vol. 7, no. 9, pp. 7-15, 2020.

[20] C. P. Kumar and L. D. Babu, "Novel text preprocessing framework for sentiment analysis," in *Smart Intelligent Computing and Applications. Springer*, Singapore, 2019.

[21] N. A. S. &. R. N. I. A. Abdullah, "Multilingual Sentiment Analysis: A Systematic Literature Review," *Pertanika Journal of Science & Technology*, vol. 29, no. 1, 2021.

[22] M. Naili and A. H. Chaibi, "Comparative Study of Arabic Stemming Algorithms for TopicIdentification," in *23rd International Conference on Knowledge-Based and Intelligent Information & EngineeringSystems*, 2019.

[23] J. Jumadi, D. S. Maylawati, L. D. Pratiwi and M. A. Ramdhani, "Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process," *IOP Conference Series: Materials Science and Engineering,* vol. 1098, no. 3, 2021.

[24] A. Roy, S. Sarkar and H. Borkakoty, "A Lemmatizer Tool for Assamese Language," in *International Conference on Computational Intelligence, Communications, and Business Analytics. Springer*, Singapore, 2018.

[25] N. Wayessa and S. Abas, "Multi-Class Sentiment Analysis from Afaan Oromo Text Based On Supervised Machine Learning Approaches," *International Journal of Research Studies in Science, Engineering and Technology,* vol. 7, no. 7, pp. 10-18, 2020.

[26] A. Konate and R. Du, "Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques," *Wuhan University Journal of Natural Sciences*, vol. 23, no. 3, pp. 237-243, 2018.

[27] G. Kaur and K. Kaur, "Sentiment detection from Punjabi text using support vector machine," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 5, no. 6, pp. 39-46, 2017.

[28] O. B. Franky and K. Veselovská, "Resources for Indonesian sentiment analysis," *The Prague Bulletin of Mathematical Linguistics*, vol. 103, no. 1, pp. 21-41, 2015.

[29] H. D. L. &. S. E. Werbin-Ofir, "Beyond majority: Label ranking ensembles based on voting rules," *Expert Systems with Applications*, vol. 136, pp. 50-61, 2019.

[30] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25-29, 2018.

[31] R. Choudhary and H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," in *2017 International Conference on Machine Learning and Data Science (MLDS). IEEE*, 2017.

[32] L. Xing, J. He, Y. Li, Y. Wu, J. Yuan and X. Gu, "Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza," *Accident Analysis & Prevention*, vol. 135, 2020.

[33] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

[34] A. Al-Anazi and . I. D. Gates, "A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs," *Engineering Geology, ELSEVIER*, vol. 114, pp. 267-277, 2010.

[35] I. M. K. Karo, A. Khosuri and R. Setiawan, "Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient," in *2021 International Conference on Data Science and Its Applications (ICoDSA). IEEE*, 2021.

[36] S. MEGI, "ALGORITMA STEMMING TEKS BAHASA KARO BERDASARKAN ATURAN TATA BAHASA," Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim, Riau, 2020.

[37] Risnawaty, Sutikno, M. Sembiring, L. Andriany and R. Siregar, "The Translation of Ngangkat Tulan-Tulan Texts in Karonese Society into English," *Journal of Talent Development and Excellence*, vol. 12, no. 1, pp. 352-361, 2020.

[38] B. Tarigan, R. Sofyan and R. N. Rosa, "Derivational morphology of Karonese ecolexicon," in *Seventh International Conference on Languages and Arts (ICLA 2018)*, 2019.

[39] S. B. Gurusinga, "PHONOLOGICAL DIALECT DIFFERENCES OF KARONESE LANGUAGE IN MEDAN, NORTH SUMATRA," *urnal CULTURE (Culture, Language, and Literature Review)*, vol. 7, no. 2, pp. 263-275, 2020.

[40] M. Sembiring and M. Girsang, "Translating Textual Theme in Maba Belo Selambar Dialogue of Karonese Society into English," in *The 1st Annual International Conference on Language and Literature*, 2018.

[41] M. Diallo, C. Fourati and H. Haddad, "Bambara Language Dataset for Sentiment Analysis," in *International Conference on Learning Representations (ICLR)* , 2021.

## BIOGRAPHY OF AUTHORS

**Ichwanul Muslim Karo Karo** is a PhD candidate in information technology, FSKTM, UTHM, Batu Pahat, Johor, Malaysia. His research interest includes Data mining, spatial mining, sentiment analysis and business intelligence. Besides, he is also actively involved as a data scientist and consultant for various projects in various institutions/organizations.

**Mohd Farhan Md Fudzee** is an Associate Professor at Fakulti Sains Komputer dan Teknologi Maklumat (FSKTM), University Tun Hussein Onn Malaysia, Malaysia. He obtained a PhD from Deakin University, Australia. His research interests include multimedia computing, green computing, decision-making, web technology, and e-government.

Shahreen Kasimis currently an associate professor in the Department of Security Information and Web Technology, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. Her areas of interest include bioinformatics, soft computing, data mining, web and mobile application.

**Azizul Azhar Ramli** is currently an Associate Professor at Fakulti Sains Komputer dan Teknologi Maklumat (FSKTM), University Tun Hussein Onn Malaysia, Malaysia. He obtained a PhD (Management Engineering), IPS, Waseda University, Japan. His research interests include data mining, soft computing, and machine learning.